tijdschrift van het

nederlands elektronica- en radiogenootschap

deel 49 – nr. 2 – 1984

nederlands elektronica- en radiogenootschap

Nederlands Elektronica- en Radiogenootschap
Postbus 39, 2260AA Leidschendam. Gironummer 94746
t.n.v. Penningmeester NERG, Leidschendam.

HET GENOOTSCHAP

De vereniging stelt zich ten doel het wetenschappelijk onderzoek op het gebied van de elektronica en de informatietransmissie en - verwerking te bevorderen en de verbreiding en toepassing van de verworven kennis te stimuleren.

Lidmaatschap

Voor lidmaatschap wende men zich tot de secretaris. Het lidmaatschap staat open voor academisch gegradueerden en hen, wier kennis of ervaring naar het oordeel van het bestuur een vruchtbarellidmaatschap mogelijk maakt. De contributie bedraagt f1.60.-per jaar.

Studenten aan universiteiten en hogescholen komen bij gevorderde studie in aanmerking voor een junior-lidmaatschap, waarbij 50% reductie wordt verleend op de contributie. Op aanvraag kan deze reductie ook aan anderen worden verleend.

HET TIJDSCHRIFT

Het tijdschrift verschijnt zesmaal per jaar. Opgenomen worden artikelen op het gebied van de elektronica en van de telecommunicatie.

Auteurs die publicatie van hun wetenschappelijk werk in het tijdschrift wensen, wordt verzocht in een vroeg stadium kontakt op te nemen met de voorzitter van de redactie commissie.

De teksten moeten, getypt op door de redactie verstrekte tekstbladen, geheel persklaar voor de offsetdruk worden ingezonden.

De abonnementsprijs van het tijdschrift bedraagt f 60,--. Aan leden wordt het tijdschrift kosteloos toegestuurd.

Tarieven en verdere inlichtingen over advertenties worden op aanvrage verstrekt door de voorzitter van de redactiecommissie.

DE EXAMENS

De door het Genootschap ingestelde examens worden afgenomen in samenwerking met de "Vereniging tot bevordering van Elektrotechnisch Vakonderwijs in Nederland (V.E.V.)". Het betreft de examens:

a. op lager technisch niveau: "Elektronica monteur N.E.R.G.";
b. op middelbaar technisch niveau: "Middelbaar Elektronica technicus N.E.R.G.".

Voor deelname, inlichtingen omtrent exameneisen, reglement, en uitgewerkte opgaven wende men zich tot het Centraal Bureau van de V.E.V., Barneveldseweg 39, 3862 PB Nijkerk; tel. 03494 - 4844.

TOEPASSING VAN SPRAAKCODERINGS- EN HERKENNINGSSYSTEMEN

Ir. F.J. Schäffers
Philips' Telecommunicatie Industrie B.V., Hilversum

Application of Speech Coding and Recognition Systems. This paper has been presented as the introduction of the course of lectures on "Speech Coding and Speech Recognition", organized by IEEE/NERG/KIVI, Delft, January 24, 1984. The paper is discussing some problems of the application of speech coding and recognition in the past and in the future. It is also describing a partition of the different speech processing techniques, together with a systematic summary of possible applications.

INLEIDING

Dit artikel werd gepresenteerd als inleiding voor de lezingendag "Spraakcodering en Spraakherkenning", georganiseerd door IEEE/NERG/KIVI te Delft, op 24 januari 1984.
Het artikel bespreekt enkele probleemsituaties die optreden of opgetreden zijn bij toepassing van spraakcodering en herkenning, nu en in het verleden. Tevens wordt een mogelijke indeling van spraakverwerkingstechnieken gegeven, aan de hand waarvan een aantal toepassingen genoemd worden.

PROBLEEMSITUATIES BIJ TOEPASSINGEN MET SPRAAK

Spraakperceptie

Spraak is het natuurlijke communicatiemiddel dat de mens ter beschikking staat.
Reeds op 3-jarige leeftijd is de mens in staat zijn gedachten uit te drukken middels spraak.
Eerst is de vocabulaire nog klein, maar deze wordt spelenderwijs vergroot, zonder dat er veel aandacht aan geschonken wordt. Al pratend en luisterend worden spraakproductie en spraakperceptie op een hoger peil gebracht, zonder dat de spreker en de luisteraar zich bewust zijn van het complexe proces dat zich aan het afspelen is.

Met name de spraakperceptie ontwikkelt zich tot een heel hoog nivo. De kracht en subtiliteit van onze perceptie van spraak zijn enorm en worden vaak onderschat. Een aardige illustratie hiervan is het volgende. Het ontwikkelen van een goede scrambler voor spraak blijkt niet eenvoudig te zijn. (Een scrambler is een apparaat waarmee een signaal zodanig vervormd kan worden, dat het onherkenbaar wordt.) Door het gebruik van een scrambler in een telefoonverbinding wordt de spraak tijdens transmissie onverstaanbaar, met het doel afluisteren door onbevoegden onmogelijk te maken. Het blijkt nu, dat zelfs wanneer de spraak ernstig verminkt wordt, de mens toch in staat is een aanzienlijk deel van de boodschap te begrijpen.

Een andere illustratie van de kracht van de menselijke spraakperceptie is het gegeven, dat veel mensen in staat zijn tot het voeren van een zogenaamd 'cocktail-party-gesprek'. Hierbij wordt deelgenomen aan een gesprek, terwijl tevens naar een conversatie van anderen geluisterd wordt: een situatie die veel voorkomt op recepties en cocktailparties.

Oftewel, onze spraakperceptie vormt zeker meer dan een eenvoudige ontvanger van een acoustisch signaal, bestaande uit klanken, woorden of zinnen.

Telefoonkwaliteit

Aan onze goed ontwikkelde spraakperceptie moet ook een deel van het succes van de telefonie toegeschreven worden. De transmissiecircuits van de eerste telefoniesystemen werden gekenmerkt door een sterke lineaire en niet-lineaire vervorming. En nog, wanneer vandaag de dag een opname van een telefoongesprek beluisterd wordt, en men let op de kwaliteit van het signaal, en niet op dat wat er gezegd wordt, dan staat men versteld van de hoeveelheid ruis, brom, kraak- en knispergeluiden die men hoort. Daarnaast zorgt de koolmicrofoon voor vervorming van de klankkleur van de stem. Echter, wanneer meer gelet wordt op de conversatie zelf, op de inhoud van het gesprek, dan valt het best wel mee met die slechte kwaliteit. Het komt eigenlijk maar heel weinig voor, dat de kwaliteit zo storend slecht is dat men besluit te verbreken en de verbinding opnieuw op te bouwen, in de hoop een betere lijn te krijgen. Met een minder goed ontwikkelde spraakperceptie zouden veel meer telefoongesprekken mislukt zijn, wat het succes van de telefonie zeker in de weg had gestaan.

Toenemend gebruik van digitale transmissiecircuits, en in de iets verder gelegen toekomst de invoering van digitale schakelcentrales en abonneeaansluitingen zullen de geluidskwaliteit enorm verbeteren. Het grote technische voordeel van digitale telefonie tegenover analoge telefonie is, dat de invloed van ruis en vervorming sterk is teruggebracht, waardoor de kwaliteit niet te lijden heeft van schakelen en

transmissie. Maar het zal nog vele jaren duren voor het telefoonnet grotendeels gedigitaliseerd zal zijn.

Bediening van de telefoon

Een ander aspect van telefonie is de bediening van het systeem. De huidige generatie en enkele generaties daarvoor, zijn opgegroeid met het gegeven dat iedere telefoonaansluiting een nummer heeft, dat met behulp van een kiesschijf of, de laatste jaren, met druktoetsen ingevoerd moet worden. Ik persoonlijk ken niemand die daar onoverkomelijke problemen mee heeft. Dit is echter niet altijd zo geweest. In de begintijd van de automatische telefonie, bij de invoering van het interlokaal en internationaal telefoneren, hadden veel gebruikers problemen met het invoeren van de (vaak lange) telefoonnummers, hetgeen als zeer gebruiksonvriendelijk werd ervaren. Dit in vergelijking met de situatie die tot dan toe bestond: het kiezen op naam via een telefoniste. De huidige generatie, die de luxe van het kiezen op naam nooit gekend heeft, beschouwt mijns inziens het kiezen op nummer als een noodzakelijk kwaad. (Een afwijkende situatie bestaat in de Verenigde Staten, waar de operator nooit helemaal is verdwenen, en die, indien gewenst, ogenblikkelijk kan bijspringen.)

In het nuvolgende deel worden twee probleemsituaties beschreven welke te maken hebben met de bediening van het telefoniesysteem. De eerste situatie betreft de bedrijfstelefonie. Moderne bedrijfstelefoniecentrales bieden vele faciliteiten, zoals 'auto-ringback' (automatisch terugbellen bij bezet), 'follow-me', niet storen, vergaderschakelingen, enzovoorts, die worden ingeschakeld door het intoetsen van een numerieke code. Deze code lijkt op een kort telefoonnummer, waarbij vaak gebruik gemaakt wordt van het hekje # en de ster * zoals deze op het druktoetsklavier van moderne telefoontoestellen voorkomen. Dus bijvoorbeeld: 'niet storen' kies ik aan door in te toetsen '* 58'. Wat blijkt nu in de praktijk?

Van de vaak tientallen faciliteiten die zo'n moderne bedrijfstelefooncentrale biedt, worden er meestal maar enkele frekwent gebruikt, en de andere bijna nooit, of slechts door een kleine groep.

Waarom?

Welnu, het blijkt dat men de procedures te moeilijk vindt om te onthouden, de numerieke codes teveel op elkaar vindt lijken, kortom, men vindt de faciliteiten niet erg vriendelijk in het gebruik. Objectief beschouwd zijn de numerieke codes korter dan de meeste telefoonnummers, en daardoor eerder gemakkelijker dan moeilijker te onthouden dan deze telefoonnummers. De codes worden echter minder aanvaard.

De tweede probleemsituatie betreft de huis-, tuin- en keukentelefonie. Moderne telefooncentrales voor openbare telefonie, zoals de gewone telefonie genoemd wordt, kunnen ons, technisch gesproken, alle diensten en faciliteiten bieden die we kunnen bedenken. De Nederlandse PTT heeft in de afgelopen jaren, met veldproeven in Amsterdam en Heerenveen, geëxperimenteerd met een aantal nieuwe diensten zoals een wekdienst, verkort kiezen, afwezigheidsmelding, niet storen, kostenopgave en herhaling van het laatst gekozen nummer. De diensten werden, net zoals bij de bedrijfstelefooncentrales bestuurd met numerieke codes, in te voeren via de druktoetsen van het telefoontoestel. Om de abonnees te helpen was het mogelijk beknopte gesproken begeleiding te geven. Ook andere Europese PTT's hebben dergelijke experimenten uitgevoerd. Bij de evaluatie van de experimenten is gebleken dat niet alle nieuwe diensten in alle landen even positief beoordeeld werden. In Nederland vond men de procedures lastig, en de numerieke codes te willekeurig gekozen. De gebruiksaanwijzing, welke onmisbaar bleef bij het gebruik, zag er door de vele codes 'te technisch' uit, wat een drempeleffekt veroorzaakte. Dus ook hier de gebruiksonvriendelijkheid welke het gebruik van de diensten in de weg staat. Voorts werden ook vraagtekens gezet bij de wenselijkheid van sommige nieuwe diensten. Of vond met de dienst wel aardig of nuttig, maar niet meer als er extra voor betaald diende te worden. Maar, het gebruik van gesproken begeleiding werd over het algemeen positief beoordeeld.

Kan de bediening gebruiksvriendelijker?

Wat doen we nu met al die technische mogelijkheden? We moeten natuurlijk voorkomen om elegante oplossingen te bedenken voor niet bestaande problemen of problemen die voor maar weinig mensen relevant zijn. Sommige diensten lijken echt wel nuttig, als het gebruik maar wat makkelijker zou zijn.

Voor het probleem van de gebruiksvriendelijkheid dienen zich twee mogelijke oplossingen aan. De eerste mogelijkheid is het gebruik van een speciaal telefoontoestel, waarop naast de 12 druktoetsen voor het kiezen een aantal extra druktoetsen aanwezig is voor de besturing van nieuwe diensten. Het mooist is natuurlijk als deze extra druktoetsen door de gebruiker te programmeren zijn voor de diensten die hij het meest gebruikt. Gesproken instructies of begeleiding zorgen voor de rest. Een aantrekkelijke oplossing, met het nadeel dat er een nieuw, waarschijnlijk niet goedkoop telefoontoestel voor aangeschaft dient te worden.

De tweede mogelijkheid is het gebruik van de stem om besturingscommando's te geven. Dus het inspreken van 'wekdienst' als we ons willen laten wekken. Dit is zeker gemakkelijker te onthouden dan een nummer uit een hele reeks. Met een gesproken begeleiding vanuit de centrale ontstaat op deze manier een gesproken dialoog tussen de gebruiker en de machine. Dit wordt door sommigen als de ideale oplossing gezien: de stem

is immers een zeer effectief middel om informatie uit te wisselen. Honderdtwintig woorden per minuut is een gemiddelde spreeksnelheid voor de meeste mensen, die met typen tot ongeveer twintig woorden per minuut komen. Daarbij komt dat je je stem altijd bij je hebt, waar je ook bent. Anderen daarentegen voelen het als onnatuurlijk om tegen een machine te praten en geven de voorkeur aan druktoetsen. Een groot voordel van de oplossing met gesproken commando's is in ieder geval dat je geen nieuw telefoontoestel hoeft te kopen voor de nieuwe diensten.

Gesproken commando's: een haalbare oplossing?

Bij de laatste oplossing, het gebruik van gesproken commando's zijn nog wel een paar kritische opmerkingen te maken. Het automatisch herkennen van spraak via de telefoon is 'a hell of a job', zelfs wanneer de te herkennen vocabulaire vrij klein is.

Om te beginnen moet de herkenner commando's herkennen van zeer veel mensen; mensen, die wellicht met een accent spreken, een verschillende moedertaal hebben, mannelijk of vrouwelijk zijn, oud of jong, en een harde of zachte stem hebben.
Ga d'r maar aan staan!

Het probleem wordt nog wat groter als men zich realiseert, dat er altijd wel achtergrond geluid aanwezig is: kantoorgeluiden, straatlawaai, muziek, pratende mensen.
Wordt de herkenner daardoor niet misleid?
En als het toch gebeurt, is er dan een eenvoudige mogelijkheid dit te korrigeren?

Helemaal problematisch lijkt het te worden wanneer de geluidskwaliteit niet constant is en soms heel erg slecht is, ten gevolge van het bestaande telefoonnet.

Een mens is onder zulke omstandigheden nog wel in staat om een gesprek te voeren, maar die heeft dan ook een geweldig goed ontwikkeld spraakperceptievermogen. Hoe een machine onder deze omstandigheden presteert is eigenlijk nog nauwelijks bekend.

In de Verenigde Staten is onlangs een vrij groot experiment uitgevoerd door de Bell Labs, de researchlaboratoria van AT&T, waarin getracht werd via de telefoon spraak van zo'n 3000 mensen te herkennen (J.G. Wilpon, 1983). Deze mensen werden door het systeem gevraagd hun telefoonnummer op te geven, uitgesproken als losse cijfers. De herkenningsvocabulaire bestond dus uit de cijfers 0 t/m 9. Het herkenningssysteem, dat onder gesimuleerde omstandigheden dik 95% van alle ingesproken cijfers juist herkende, bleek in de praktijkproef met 47% van alle proefpersonen problemen te hebben. Van de overige 53% werd ruim 77% juist herkend. De 47% probleemgevallen waren voor een deel het gevolg van een slechte telefoonverbinding of achtergrondgeluiden. Maar ook een flink deel werd door de proefpersonen veroorzaakt, die òf niet normaal spraken (waarschijnlijk veroorzaakt door de instructie in losse cijfers te spreken, waardoor bijvoorbeeld overdreven werd gearticuleerd), òf de instructies niet begrepen en hun telefoonnummer uitspraken als twaalf-vierendertig-zesenvijftig. Door de gesproken instructie van het systeem aan de proefpersonen bij te stellen werd de laatste groep overigens aanzienlijk verkleind.

Hoe nu verder?

Moet uit dit experiment nu de conclusie getrokken worden, dat spraakherkenning via de telefoon een bijna onmogelijke opgave vormt?
Ik denk dat het nog wat vroeg is voor conclusies, want er zijn nog heel wat onbeantwoorde vragen, ook met betrekking tot het genoemde experiment.

Bijvoorbeeld, het spraakgedrag van de proefpersonen is blijkens enkele experimenten, waaronder het bovenbeschreven experiment, beïnvloedbaar. Onderzocht dient dus te worden wat de optimale wijze van instructie of begeleiding is.

Een tweede punt is in hoeverre de resultaten afhankelijk van de herkenningsvocabulaire zijn. Het is bekend dat er, zowel voor de mens als de machine, moeilijke en makkelijke herkenningsvocabulaires zijn. Denk maar aan het spellingsalfabet. Niet voor niets wordt bij spelling onder moeilijke omstandigheden gebruik gemaakt van hele woorden, zoals Alpha, Bravo, Charlie, Delta, ...

Een derde punt betreft de spraakherkenner zelf. Tot nog toe zijn de spraakherkenners eigenlijk acoustische herkenners, die een hoestbui of keelschrapen net zo serieus trachten te herkennen als een gesproken commando. De menselijke spraakherkenning maakt gebruik van fonetische informatie, evenals van informatie over de context van de te herkennen uitspraak in de ruimste zin van het woord: kennis over de spreker, diens omstandigheden en emotionele toestand, enzovoorts.
Het ligt in de lijn der verwachtingen, dat de betrouwbaarheid van automatische spraakherkenning zal toenemen door het gebruik van deze informatie in een 'intelligente' herkenner.

Een experiment als dat van Bell is van grote waarde. Deze experimenten geven veel inzicht in de problematiek van spraakherkenning, zelfs wanneer eenvoudige herkenningsapparatuur met beperkte mogelijkheden gebruikt wordt. Er valt namelijk nog heel wat te leren over de zogenaamde 'human factors', de gebruikersaspecten, zoals dialoogstructuren en dergelijke. Dat lukt eigenlijk alleen maar buiten het laboratorium in de praktijk, met niet-coöperatieve proefpersonen. Ook in Nederland bestaan plannen voor een spraakherkenningsexperiment via de telefoon. Op dit moment wordt gewerkt aan de opzet van een gezamen-

lijk experiment door de PTT en Philips.

Niet altijd zal toepassing van spraakherkenning zoveel problemen opleveren. In een aantal toepassingen zal het aantal te herkennen sprekers beperkt zijn tot één of een paar. Of kan er gebruik gemaakt worden van een goede microfoon, die geen last heeft van omgevingslawaai. Of men heeft veel vrijheid bij het samenstellen van de herkenningsvocabulaire, zodat deze optimaal gekozen kan worden. Onder zulke gunstige omstandigheden zijn vandaag de dag goede resultaten haalbaar.

Nota Bene

In deze inleiding heb ik opzettelijk veel aandacht besteed aan de problemen die optreden bij toepassing van automatische spraakherkenning bij de mens-machine communicatie. De reden hiervoor is, dat ik wat tegengas wil geven aan die personen, die denken dat het probleem van spraakherkenning wel opgelost zal worden door gebruik van krachtiger en snellere computers, of een efficiëntere zoekprocedure.

Nee, de menselijke spraakherkenning, de sterk ontwikkelde spraakperceptie, waar we de prestaties van automatische spraakherkenning aan spiegelen, is nog altijd niet goed begrepen. Een beter inzicht in het perceptieproces van de mens zal de automatische herkenning structureel kunnen verbeteren. Zolang dit inzicht ontbreekt zullen slechts marginale verbeteringen mogelijk zijn.

SPRAAKVERWERKINGSTECHNIEKEN

Een mogelijke indeling van spraakverwerkingstechnieken is de volgende:

- Spraakinvoer : . spraakcodering
  . spraakherkenning
  . sprekerherkenning
- Spraakuitvoer: . spraakdecodering
  . tekst- spraakomzetting
- Dialoog : . dialoogbesturing

Hieronder worden de verschillende technieken kort gedefiniëerd.

Spraakcodering is de omzetting van analoge spraak naar digitale code, waarbij gestreefd wordt naar behoud van kwaliteit zonder te letten op de inhoud van de spraak (dus op wât gezegd wordt).

Spraakherkenning is de herkenning van een uitspraak, door vergelijking van een digitaal patroon van de uitspraak met van tevoren gemaakte referentiepatronen van te herkennen uitspraken.

Sprekerherkenning is de verificatie van een opgeëiste identiteit, aan de hand van een toetsuitspraak.

Spraakdecodering is de omzetting van voorheen gecodeerde spraak (= digitale code) naar hoorbare analoge spraak.

Tekst-spraakomzetting is de omzetting van tekst (ASCII of fonetisch) naar analoge spraak door spraakeenheden uit een bestand aaneen te schakelen, onder toepassing van regels voor overgangen, intonatie, luidheid en tijdsduur.

Dialoogbesturing is de beheersing van het verloop van de dialoog tussen mens en machine, teneinde een efficiënte en gebruiksvriendelijke procedure te verkrijgen.

Bijv.:
- bewaking van voortgang/syntax/tijd/timing.
- detectie van oproep/verbreken (telefonie).
- aktie na herkenning/uitblijven commando/ illegaal commando/detectie verbreken/protest van gebruiker (correctie).

TOEPASSINGEN

In de nu volgende paragrafen wordt voor iedere spraakverwerkingstechniek een aantal toepassingen genoemd, zo mogelijk gegroepeerd naar aard en situatie van toepassing.

Spraakcodering en -decodering

Transmissie van spraak
- 'gewone' telefonie
- mobiele telefonie (auto-telefoon)
- militaire communicatie

Opslag van spraak
- meldingen aan telefoonabonnee (nummerwijziging, storingsmelding, beurtmelding, tijdmelding, weerbericht)
- automatisch beantwoordingsapparaat
- Voice Mail/Voice Store and Forward

'Eyes-busy' situaties
- sprekende meetinstrumenten (vliegtuig, laboratorium, microscopie)

Visueel gehandicapten
- sprekende klokken, horloges, rekenmachines
- telefonie: bedienpost voor blinde telefonisten

Telemetrie
- aflezen van meetinstrumenten via telefoon

Waarschuwingssystemen
- brandmelding (kantoren, warenhuizen, hotels)
  NB. niet alleen melding, maar ook instructie!
- inbraakalarm (stil alarm via telefoon)
- auto: sprekend dashboard
- melding van bedieningsfouten + instructie, ingebouwd in apparatuur
- halte-afroepsystemen voor openbaar vervoer

Spraakherkenning

'Eyes-busy'/'Hands-busy' situaties
- CAD-werkstations
- bediening (sprekende) instrumenten in vliegtuigen
- microscopie: besturing van apparatuur
- afregelen van apparatuur of systemen
- sorteren: bagage op luchthavens, magazijnen
- kwaliteitscontrole
- invoeren van grote hoeveelheden gegevens

Telefonie: vervanging toondruktoetskeuze
- aankiezen speciale functies van bedrijfscentrale (auto-ringback, follow-me, niet storen)
- aankiezen bijzondere diensten (storingsdienst, inlichtingen, weerbericht, wekdienst)
- aankiezen instanties (politie, brandweer)
- aankiezen afdelingen of personen op naam

Visueel gehandicapten
- telefonie: bedienpost voor blinde telefonisten.

Sprekerherkenning
- toegangscontrole tot gebouwen, gegevensbestanden, faciliteiten
- validatie van (telefonische) transacties

Tekst-spraakomzetting

Gehandicapten
- leesmachines voor blinden
- spreekmachines voor spraakgestoorden

Weergave vaak wijzigende tekst
- weerbericht
- informatie uit computerbestanden
- halte-afroepsystemen voor openbaar vervoer
- vertaalsystemen

Dialoogbesturing

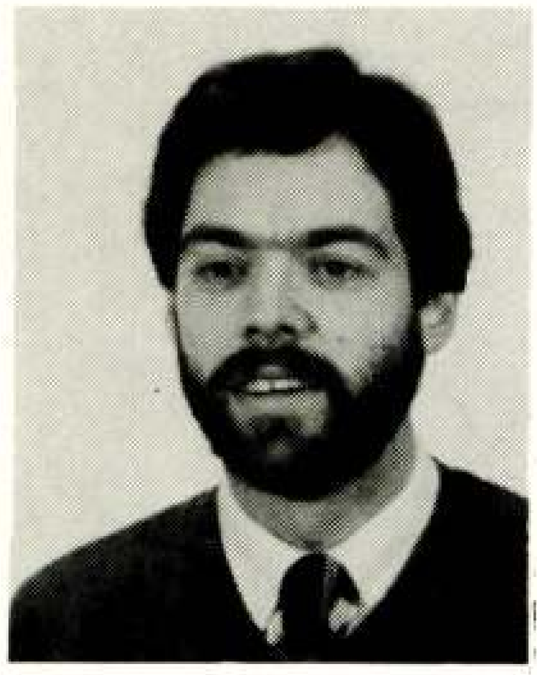
Informatie verstrekkende diensten
- verkeersinformatie
- weersinformatie
- aankomsttijden vliegtuigen
- waterstanden rivieren
- bankinformatie (mutaties, saldo)
- telefoonnummers
- koersinformatie (beurs)

Diverse toepassingen
- Voice Mail/Voice Store and Forward
- reserveren voor vliegtuigen en treinen
- telefonisch bestellen bij postorderbedrijven
- telefonisch reserveren voor evenementen (toneel, muziek, sport)

LITERATUUR

J.G. Wilpon, L.R. Rabiner, "On the Recognition of Isolated Digits From a Large Telephone Customer Population", The Bell System Technical Journal, Vol. 62, No. 7, September 1983, P 1977-2000.

Voordracht gehouden tijdens de 319e werkvergadering.



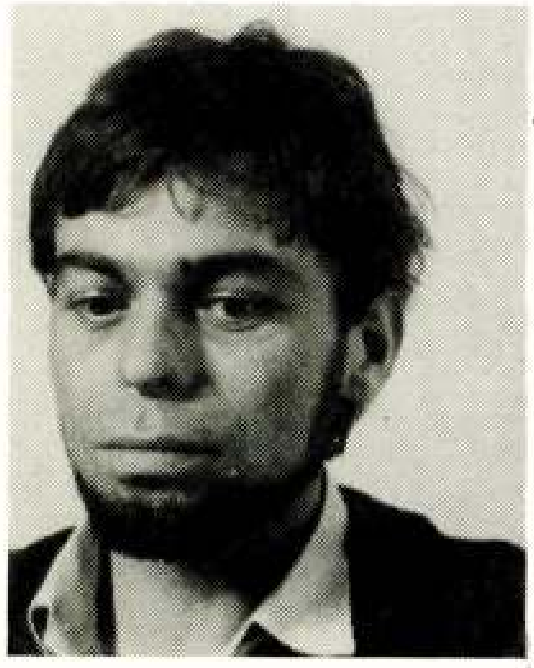
IEEE STUDENT BRANCH DELFT
IEEE BENELUX SECTION
NEDERLANDS ELEKTRONICA- EN RADIOGENOOTSCHAP
(319e werkvergadering)
SECTIE TELECOMMUNICATIETECHNIEK, KIvI

UITNODIGING

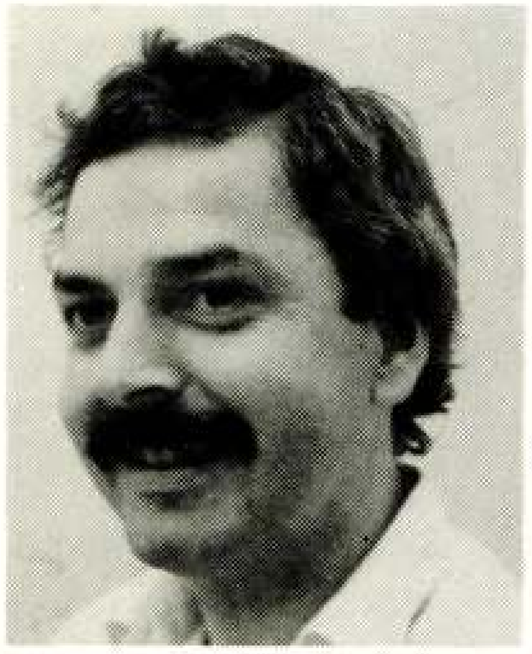
voor de lezingendag op **24 januari 1984** in **zaal A** van het **gebouw** voor **Elektrotechniek** van de **Technische Hogeschool Delft.**

THEMA: SPRAAKCODERING EN SPRAAKHERKENNING.

PROGRAMMA

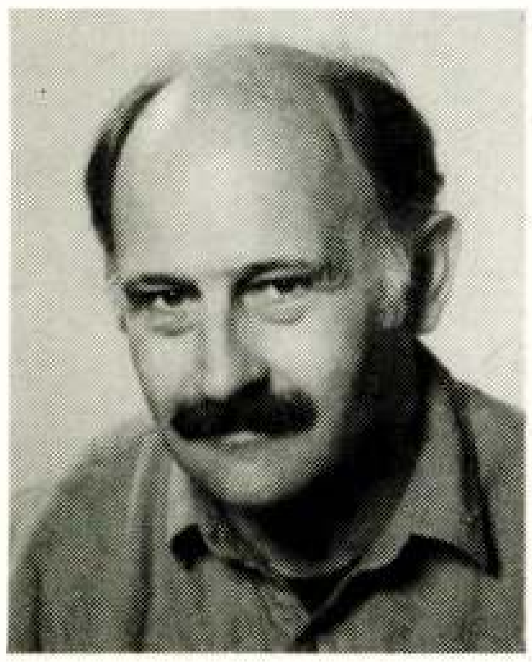


09.30 uur:	Ontvangst en koffie.
10.00 uur:	**IR. F. J. SCHÄFFERS**,(Philips Telecommunicatie Hilversum); Foto 1 TOEPASSING VAN SPRAAKCODERINGS- EN HERKENNINGSSYSTEMEN.
10.30 uur:	**DR. IR. E. F. A. DEPRETTERE**, (TH Delft, Vakgroep Netwerktheorie); Foto 2 SPRAAKCODERING: OVERZICHT EN HUIDIGE TRENDS.
11.15 uur:	Koffiepauze. Foto 3
11.30 uur:	**IR. L. J. P. VAN HEUGTEN**, (DNL-PTT, Leidschendam); SPRAAKHERKENNING: EEN OVERZICHTSVERHAAL.
12.15 uur:	Lunchpauze.
13.30 uur:	**DR. M. BOOT**, (Instituut voor toegepaste taalkunde en computer linguistiek Rijksuniversiteit Utrecht); AUTOMATISCH VAN TEKST NAAR SPRAAK.
14.15 uur:	**IR. L. F. WILLEMS**, (Instituut voor Perceptie Onderzoek, Eindhoven); Foto 4 SPRAAKSYNTHESE: STAND VAN ZAKEN EN TOEKOMST.
15.00 uur:	Theepauze.
15.30 uur:	**G. J. BOSSCHA**, (Philips Nat. Lab., Eindhoven); Foto 5 HARDWARE VOOR SPRAAKCODERINGSSYSTEMEN, GENERAAL PURPOSE EN SPECIAL PURPOSE CHIPS.
16.15 uur:	**H. J. M. STEENEKEN**, (Instituut voor Zintuigfysiologie, TNO Soesterberg); Foto 6 EVALUATIE VAN SPRAAKPRODUKTIE EN SPRAAKHERKENNINGSSYSTEMEN.
17.00 uur:	Sluiting.



Aanmelding dient te geschieden door inzending van de aangehechte kaart, gefrankeerd met een postzegel van 50 cent, alsmede overmaking van de verschuldigde kosten op postrekening 4314755 van de IEEE Student Branch Delft te Delft onder vermelding van "Spraak".
De aanmelding is alleen geldig indien de aanmeldingskaart en overschrijving zijn ontvangen vóór 17 januari 1984.

De kosten voor leden van IEEE, NERG en KIvI bedragen *f* 10,— en voor introducée's *f* 25,— per deelnemer, inclusief proceedings. De lunch is niet inbegrepen, maar deelnemers kunnen gebruik maken van kantine faciliteiten van de afdeling Elektrotechniek. Studenten hebben gratis toegang tot de lezingen.

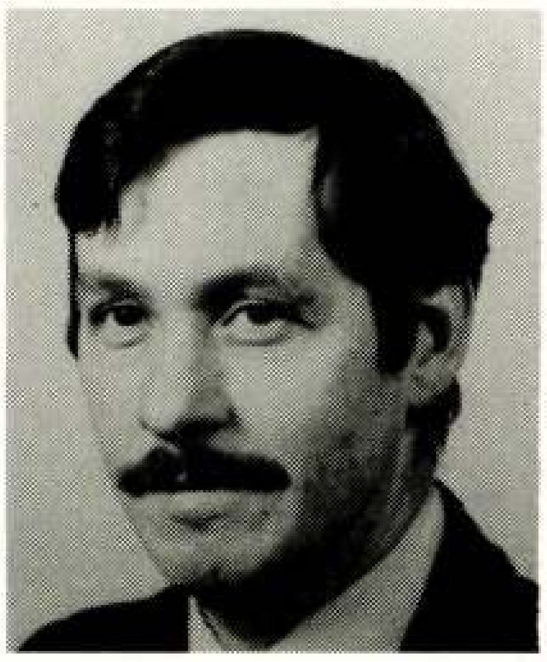


De TH Delft (Elektrotechniek) is bereikbaar met het openbaar vervoer.
Treinreizigers dienen uit te stappen bij station Delft. Vanaf dit station vertrekt elk halfuur bus 63 richting TH-wijk, uitstappen bij halte Elektro.
De TH Delft ligt halverwege de rijksweg A13 Rotterdam – Den Haag. Uit beide richtingen moet men de afslag Delft Zuid nemen. Vandaar is met ANWB borden de weg naar de TH-wijk aangegeven. Het gebouw van Elektrotechniek ligt aan de Mekelweg 4 en is makkelijk herkenbaar aan de rood/blauwe hoogbouw.

Namens de samenwerkende verenigingen,
IR. P. KROON.
Tel. 015 - 78 62 89

Delft, december 1983.





COMPRESSION AND QUANTIZATION OF SPEECH

Ed.F.Deprettere and P.Kroon
Delft University of Technology
Department of Electrical Engineering
Mekelweg 4, 2628 CD Delft, The Netherlands

This paper gives an overview of the methods and techniques used for efficient storage and/or transmission of speech signals. For all these methods the primary goal is to obtain high coding efficiency with the least possible distortion. Knowledge of speech characteristics, speech perception processes, and usage of speech models are crucial factors in the design process of speech coders. Sophisticated techniques using these factors lead to high quality coders and some of these techniques have already been implemented in real time.

INLEIDING

Spraak kan worden gerepresenteerd door informatie dragende symbolen, en het minimaliseren van de hoeveelheid van zulke symbolen is het wezenlijke van spraakcompressie. De zin ervan ligt in de ermee samenhangende reductie van de grootte van het geheugen of de bandbreedte voor respectievelijk de opslag en de transmissie van de representatie symbolen. Vrijwel zonder uitzondering wordt een regeneratie van het oorspronkelijk signaal beoogt. Een zekere mate van distorsie is daarbij onvermijdelijk. Het rendement van een compressie systeem hangt af van het compressie gehalte, de mate van distorsie, de storingsgevoeligheid en de implementatie complexiteit. Aan de complexiteit kunnen (alsnog) strikte beperkingen worden opgelegd in gevallen dat een implementatie in reele tijd gewenst is, zoals bij transmissie van spraaksignalen. Bij opslag is dit vaak alleen het geval bij het terugwinnen van de spraak. Vrijwel alle moderne compressie systemen, hebben gedigitaliseerde spraak als ingangssignalen. Dit verhoogt de handelbaarheid, de willekeurige toegangkelijkheid, de bereikbare signaal-ruis verhouding en de mogelijkheid tot foutloze transmissie. Maar tegelijk resulteert conversie van analoge naar digitale signalen in een grotere bandbreedte voor transmissie, wat op zichzelf al om compressie kan vragen. Met moderne compressie algoritmen wordt gestreefd naar een verdergaande reductie van de bandbreedte, bijv. ten behoeve van radio en satelliet communicatie, of voor geintegreerde spraak/data communicatie netwerken. Het spreekt bijna vanzelf dat een compressie systeem ontwerp beinvloed zal worden door o.a. de beschikbare opslag of transmissie capaciteit (de zgn. bit frekwentie uitgedrukt in kbit/s) en de, overigens moelijk meetbare, kwaliteit van de teruggewonnen spraak. Men verwacht dat de laatste factor afhankelijk zal zijn van de eerste, maar een universeel systeem met een geleidelijk verband tussen kwaliteit en bit frekwentie bestaat (nog) niet.

Traditionele compressie methoden kunnen ingedeeld worden in twee categorieen. De eerste categorie bevat die methoden die de redundantie in het signaal uitbuiten. Redundantie heeft te maken met voorspelbaarheid/willekeur en met dynamiek. Volledig voorspelbare signalen dragen slechts een beperkte hoeveelheid informatie. Ze kunnen geconstrueerd worden met behulp van een voorspellingsregel en een reeks van startwaarden: samen de werkelijke informatie dragers. Regeneratie van volledig willekeurige signalen is slechts mogelijk door ze als zodanig op te slaan. Praktische signalen zijn noch het een, noch het ander, en vele compressie algoritmen zijn erop gericht een gegeven signaal s(n) om te zetten in een ander, zeg r(n), dat niet redundant en efficienter codeerbaar is en waaruit het oorspronkelijk signaal teruggewonnen kan worden. Het zijn causale technieken zoals gewone en logaritmische puls code modulatie (PCM en log-PCM), differentiele puls code modulatie (DPCM) en andere predictie-coderings methoden.

In de tweede categorie is compressie het gevolg van niet-momentane, energie behoudende en omkeerbare blok transformaties, waarmee de informatie geconcentreerd wordt in een beperkt aantal transformatie componenten. Voorbeelden zijn tal van orthogonale transformatie coderingen waaronder de Fourier en de Cosinus transformatie. In alle gevallen wordt kwantisatie toegepast en dus distorsie ingevoerd. Efficiente compressie technieken streven ernaar deze distorsie te minimaliseren.

Methoden in beide categorieen worden golfvormcoders genoemd omdat ze de reconstructie van de signaalvorm op het oog hebben. Ze zijn bovendien zeer robuust en algemeen toepasbaar en worden evengoed gebruikt voor beeldcodering, waar ook wel mengvormen worden aangetroffen.

Spraak heeft evenwel zeer specifieke kenmerken, en het uitbuiten daarvan, in bijzondere methoden, is de aangewezen manier om hogere compressie verhoudingen te bereiken. Dit zal het onderwerp van discussie zijn in de volgende paragrafen waarin aan bod zullen komen de adaptieve golfvormcoders, de vocoders (modelcodering), en tussen- en mengvormen. Men zal er echter op bedacht moeten zijn, dat gespecialiseerde algoritmen in het algemeen ook fout gevoeliger zijn.

Tot slot van deze inleiding volgt nog een indicatie van wat haalbaar is met de algemene golfvormcoders, in het bijzonder de eerst genoemde tijd-domein golfvormcoders. Als referentie nemen we de zgn. u-wet of A-wet CODEC, d.i. digitale logaritmische pulscode modulatie spraak (log-PCM) van zgn. 'telefonie' kwaliteit. Dus, zie fig. 1., analoge spraak wordt in bandbreedte begrensd tot 4 kHz en in de tijd

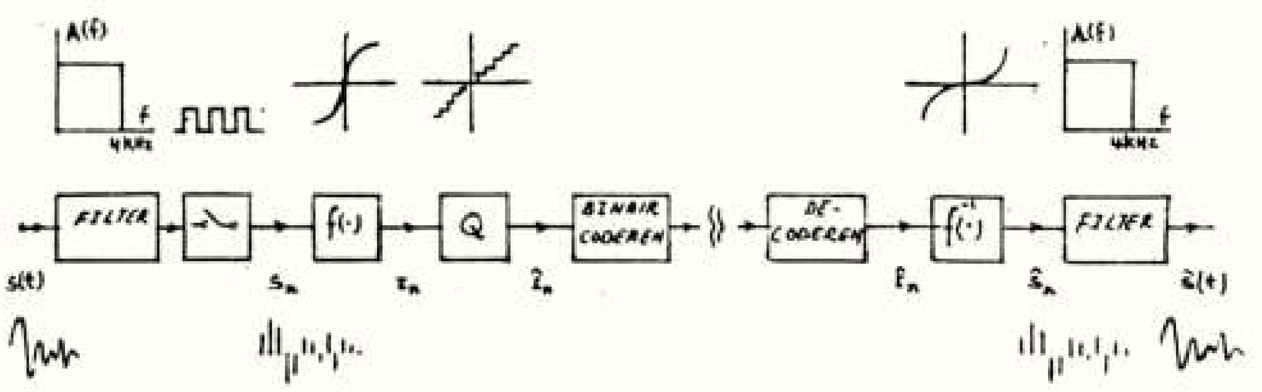


Fig. 1. Elementaire Compander Codering

distorsievrij gekwantiseerd tot 8.000 monsters per seconde. De bemonsterde spraak wordt dan logaritmisch gecomprimeerd, in amplitude uniform gekwantiseerd met 8 bits/monster en binair gecodeerd. De gedecodeerde spraak wordt exponentieel geexpandeerd en ten slotte weer analoog gemaakt.

De amplitude kwantisering introduceert distorsie, de zgn. kwantisatie ruis. Met log-PCM wordt een bit frekwentie van 64 kbit/s bereikt en een signaal-ruis verhouding die zonder de 'compander', dus bij gewone puls code modulatie (PCM) een 12-bits kwantisering zou vergen.

De meer representatieve methode uit categorie 1 is DPCM. Differentiele puls code modulatie is schematisch getoond in fig. 2.

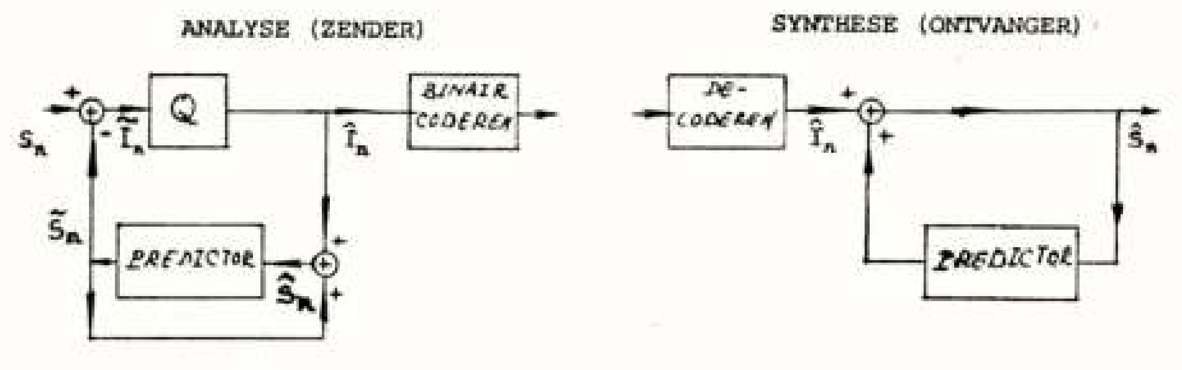


Fig. 2. DPCM-Coder

Daarin is $\hat{s}(n)$ de reproductie van $s(n)$. De zender voorspeller onthoudt vorige reproductie monsters $\hat{s}(n-1)$, $\hat{s}(n-2)$,... om de waarde van $s(n)$ te voorspellen, wat $\tilde{s}(n)$ oplevert. De voorspellingsfout $s(n) - \tilde{s}(n) = \tilde{r}(n)$ wordt gekwantiseerd tot $\hat{r}(n)$. De (identieke en omgekeerd opererende) ontvanger voorspeller reproduceert $\hat{s}(n)$ uit $\hat{r}(n)$ en $\hat{s}(n-1)$, $\hat{s}(n-2)$,... . De winst t.o.v. gewone PCM, bij overigens gelijke kwantiseringsruis, is evenredig met de logaritme van de verhouding van de vermogens van $s(n)$ en $\tilde{r}(n)$. Deze zijn gelijk als $s(n)$ volledig willekeurig, dus niet voorspelbaar is. De winst neemt dus toe met de voorspelbaarheid van $s(n)$. In DPCM is de voorspeller constant en optimaal voor lange termijn voorspelling. Als de kwantiseringsstapgrootte regelmatig wordt aangepast (ADPCM) voor optimale dynamiek compressie is een bitfrekwentie van 32 kbit/s goed haalbaar. Deze methode zal vermoelijk worden aanbevolen als nieuwe standaard voor het publiek telefoon verkeer.

Maar 32 kbit/s is niet het einde. Rekening houden met de wijze waarop spraak geproduceerd wordt leidt tot korte-termijn voorspelling, en tot modelvorming: dit wil zeggen tot een beschrijving van de bron van informatie en het golfvormingsproces. Rekening houden ook met het waarnemingsmechanisme leidt tot aanwijzingen omtrent de te volgen kwantiseringsstrategie, in het bijzonder tot maskering van de kwantisatieruis. De volgende paragrafen zijn gewijd aan compressie methoden die zulke spraakspecifieke kenmerken uitbuiten.

ADAPTIEVE GOLFVORMCODERING

DPCM gebruikt een constante voorspeller die optimaal is voor over lange perioden waargenomen spraaksignalen. Hij is kort, dit wil zeggen, gebruikt slechts 1 à 2 vorige monsters om het huidige te voorspellen. Meer heeft geen zin. Maar spraak is fundamenteel niet stationair: er is een grote variatie in amplituden; stemhebbende klanken, stemloze klanken en stilte perioden volgen elkaar voortdurend en ongeregeld op, zodat lange termijn waarneming voorbij gaat aan de werkelijke redundantie. Aan de andere kant is het fysische spraak produktie systeem mechanisch traag, als gevolg waarvan spraak nagenoeg stationair is in intervallen van 10 à 20 ms (milliseconden). Binnen zulke intervallen heeft spraak een zeer gedetailleerd spectrum en kan de redundantie aanzienlijk zijn. Anders gezegd, een over 10 à 20 ms constante voorspeller kan op basis van een tiental vorige monsters het signaal binnen het interval met een geringe fout voorspellen. In fig. 3 wordt een lang spraaksegment en 2 moment opnamen, respectievelijk een stemhebbend en een stemloos fragment getoond. Fig. 4 toont de beide fragment spectra.

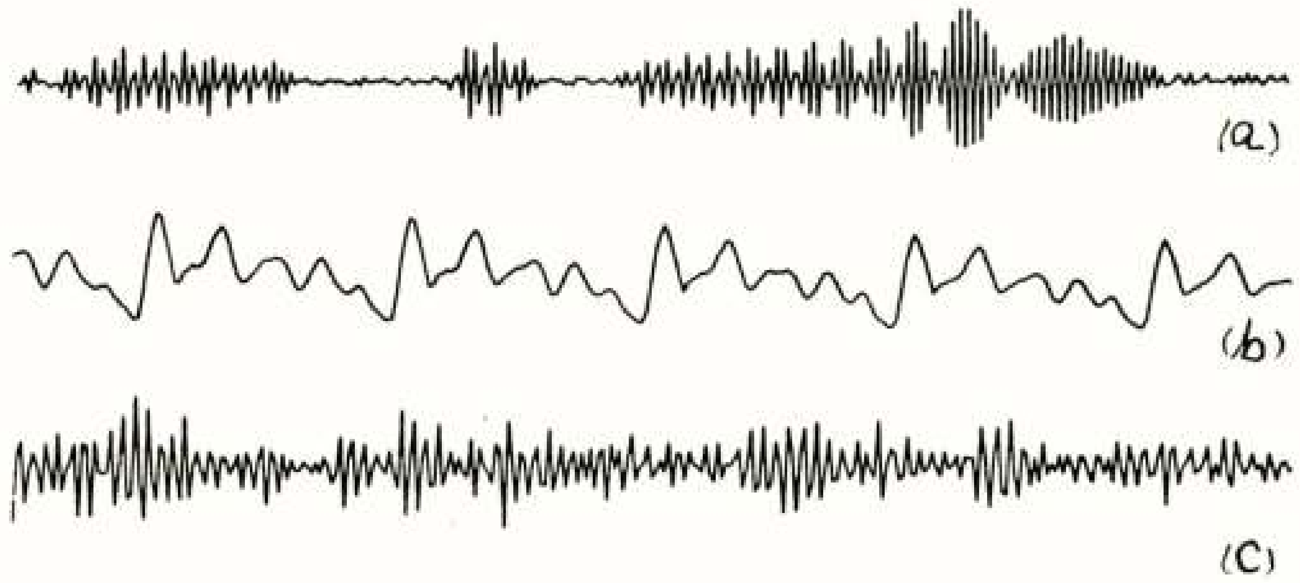


Fig. 3. Spraak in beeld. Een lang segment (a), en 2 detailopnamen: stemhebbend (b) en stemloos (c).

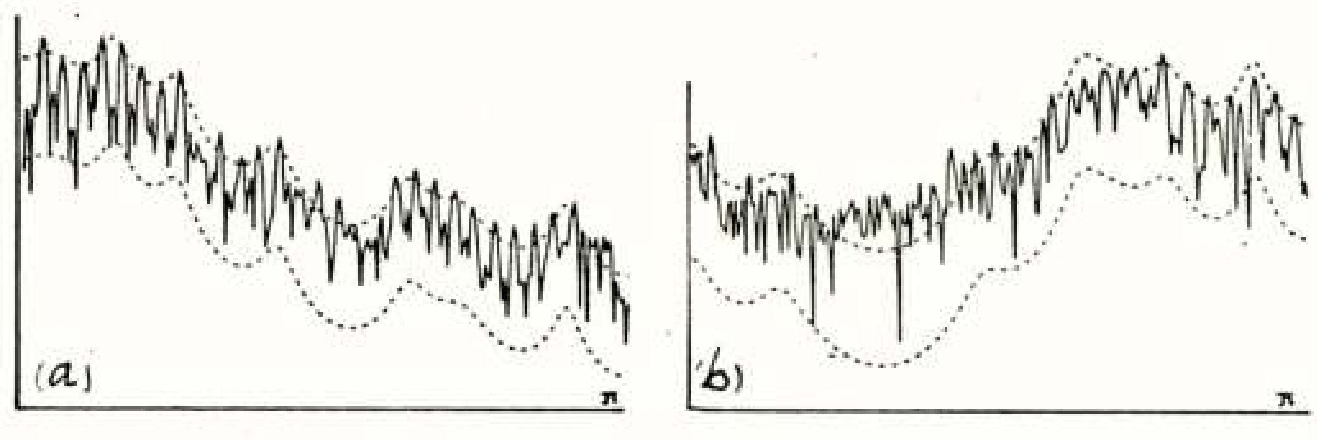


Fig. 4. Spectra van de fragmenten uit fig. 3. Stemhebbend (a) en stemloos (b).

De stippellijnen zijn de zgn. spectrale omhullenden waarop we verder nog terug zullen komen. De gearceerde gebieden geven aan waarbinnen het vermogen van eventueel aan de spraak toegevoegde ruis of vervorming vallen mag wil het oor er niet door gehinderd worden. Moderne coderingsalgoritmen maken daarvan gebruik door de kwantisatie ruis zodanig spectraal te vormen (te kleuren) dat deze binnen dit zgn. maskeringsgebied valt. Met locale voorspelling en ruiskleuring kan de DPCM compressie verhouding worden verdubbeld tot bitfrekwenties van 16 kbit/s. Algoritmen die dit realiseren, ten koste van een verhoogde complexiteit, worden predictie adaptieve coderingsmethoden genoemd (APC). Een prototype is getoond in fig. 5.

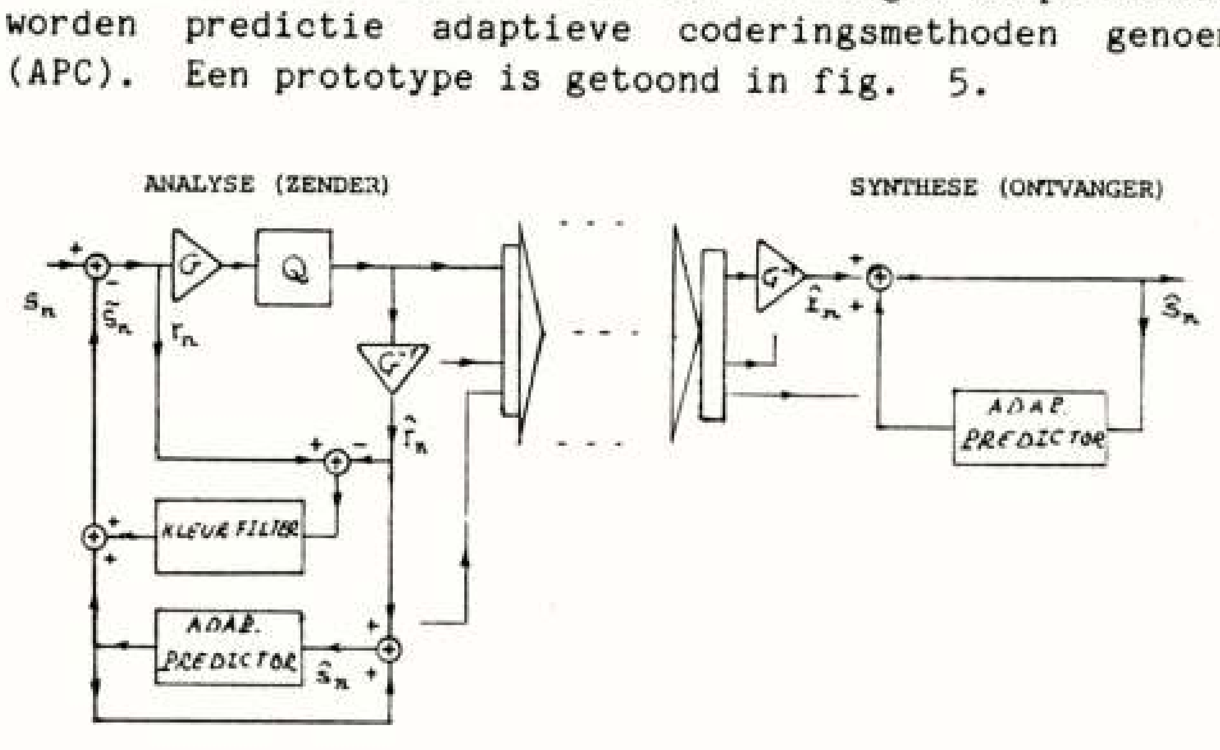


Fig. 5. APC systeem.

Daarin is de voorspeller 10 à 16 monsters lang en wordt om de 10 à 20 ms optimaal bijgesteld. Het kleuringsfilter zorgt voor maskering van de kwantisatie ruis. Het kwantiseren is net als bij ADPCM adaptief doordat de uniforme karakteristiek Q toegepast wordt op het segmentsgewijs optimaal geschaalde (G) residu signaal $r(n)$. Merk op dat ten minste 2 à 3 van de 16 kbit code nodig is voor codering van de voorspellingsparameters en de schaalfactor. APC heeft

de adaptieve transformatie coder (ATC) uit de 2e categorie coderingsmethoden tot tegenhanger. In ATC algoritmen wordt het aantal toegekende code bits per transformatie component dynamisch gevarieerd over de verschillende componenten. Een tussenvorm is de zgn. sub-band codering (SBC) die in feite een mengvorm is uit de 2 in de inleiding genoemde categorieeñ coderingsmethoden. In deze methode, zie fig. 6, wordt spraak geanalyseerd met een (constante) filterbank.

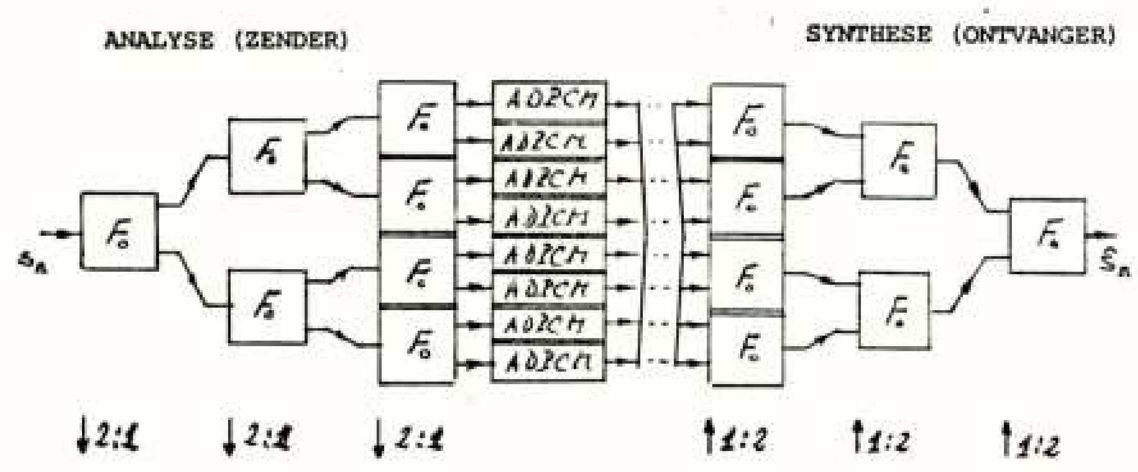


Fig. 6. Sub-Band Coder systeem.

De signalen in de verschillende sub-banden worden dan in bemonsteringstempo verlaagd en volgens de ADPCM techniek gekwantiseerd. Korte termijn details worden tot op zekere hoogte vastgelegd, en ook ruismaskering is tot op zekere hoogte gegarandeerd omdat het distorsie vermogen in elke band evenredig is met het signaalvermogen. SBC is implementatie technisch uiterst aantrekkelijk, kwalitatief echter is een bitfrekwentie lager dan 24 kbit/s niet aanvaardbaar.

Met deze methoden is de grens van de aanvaardbare compressie bereikt. De kwaliteit van de teruggewonnen spraak degradeert op haast discontinue wijze wanneer gecodeerd in het gebied onder de 16 kbit/s. Om deze grens te doorbreken is modelvorming nodig. Deze benadering wordt in de nu volgende paragraaf toegelicht.

VOCODERS

Het spraakproduktie proces kan worden gemodelleerd als getoond in fig. 7a. Dit model bestaat uit een geluidsbron en een golfvormende kanaal.

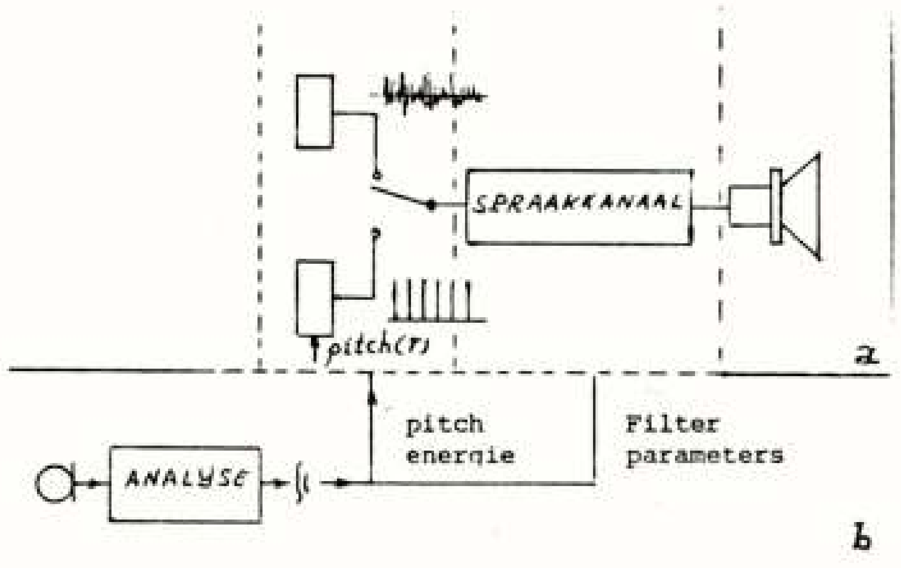


Fig. 7. Model voor het spraakproduktiesysteem (a) en analyse voor codering op basis van dit model (b).

De 2 bronsignalen reflecteren de (overgesimplificeerde) opvatting, dat spraak of stemhebbend of stemloos is en gegenereerd wordt door respectievelijk stemband trillingen of turbulente luchtwervelingen. De excitatie generator levert dus een pulstrein met pulsafstand T (de toonhoogte bepalende pitch) voor stemhebbende spraak, of een witte ruis voor stemloze spraak. Het spraakkanaal is een akoestische pijp met een vijftal eigenresonanties. Het zijn de zgn. formanten die zich duidelijk manifesteren als hobbels in de spraak spectra omhullenden getoond in fig. 4 (stippellijn). De bronsignalen zijn verantwoordelijk voor de detailstructuur in deze spectra. Een vocoder is een spraakcoderingssysteem waarin het model van fig. 7a gebruikt wordt voor de produktie van synthetische spraak. Het spraakkanaal wordt met een tijd varierend filter gerealiseerd. De vereiste informatie bestaat uit de filter parameters, de beslissing stemhebbend/stemloos/stilte, het bronvermogen en, indien stemhebbend, de pitch T. Deze informatie wordt bepaald en gecodeerd door de zender die daartoe gesegmenteerde natuurlijke spraak analyseert, zie fig. 7b. Wegens het korte termijn stationaire karakter van spraak zijn de segmenten weer 10 à 20 ms lang. Het meten van de pitch in stemhebbende spraak kan worden uitgevoerd op zowel het spraaksignaal zelf als op een frekwentie transformatie daarvan. Ook voor het bepalen van de kanaalfilter parameters zijn tijddomein en frekwentiedomein technieken mogelijk. Zo maakt de kanaalvocoder gebruik van een filterbank. De DFT-vocoder gebruikt een getrapte benadering van de spraaksegment spectra. De LPC-vocoder gebruikt lineaire predictie technieken om deze spectra met een zgn. autoregressief filter omhullend te benaderen (stippellijn in fig. 4). Fig. 8 toont het spectrum van zo'n autoregressief filter, van een puls trein (pitch T = $1/f_o$), en van het bijbehorende synthetisch spraakfragment.

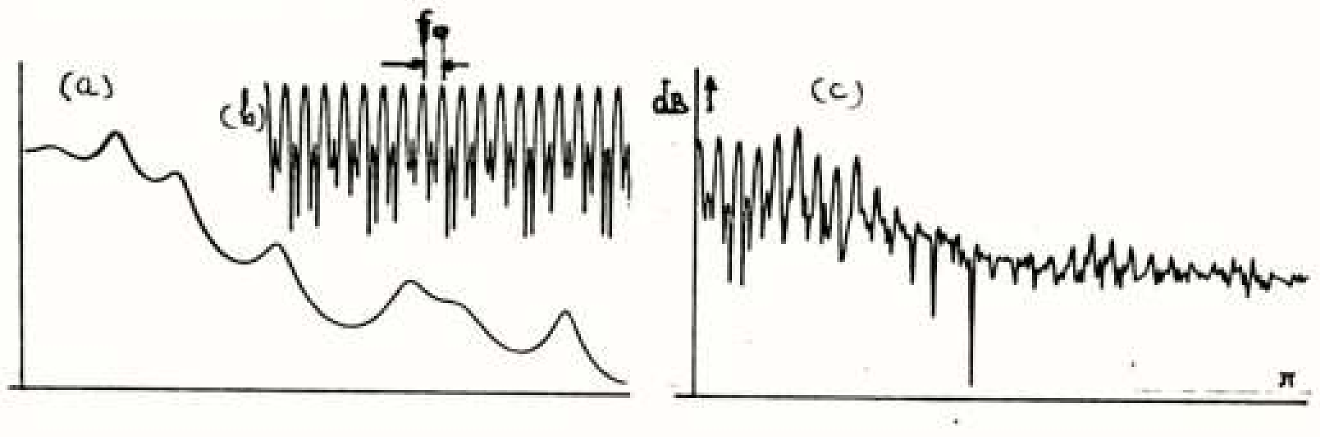


Fig. 8. Spectra van het spraakkanaal (a), van een pulstrein (b) en van het bijbehorende spraak segment (c).

Het originele segment is dit van fig. 3b/fig. 4a. Vocoder compressie is aanzienlijk. Bit frekwenties onder de 4 kbit/s zijn bereikbaar. De teruggewonnen spraak mist echter de natuurlijkheid van golfvormcoder spraak, al is de verstaanbaarheid goed tot zeer goed. Hogere bitfrekwenties helpen daarbij niet omdat de beperkingen primair het gevolg zijn van een al te simpele voorstelling van zaken. Aan de andere kant kan, met een beperkt kwaliteitsverlies, gecomprimeerd worden tot tegen de 0.4 kbit/s. Daartoe zijn echter zeer geavanceerde, complexe en vertragende coderingstechnieken vereist. Zoals vectorkwantisatie, waarbij de optimale filter parameters van elk spraaksegment bloksgewijs gecodeerd worden door ze te vervangen door een best passende parameter vector uit een verzameling van representatieve en experimenteel bepaalde filter parameter waarden. Zelfs de taalstructuur kan in de analyse betrokken worden waardoor meerdere segmenten en bloc gecodeerd kunnen worden. Maar dit leidt ons te ver af van de taak kwaliteit te garanderen bij bitfrekwenties tussen de 4kbit/s en de 16 kbit/s. De compresssie methoden in de nu volgende paragraaf richten zich op dit gebied.

BRONCODERING

Het blijkt dat het optimale synthese filter van de LPC vocoder niets anders is dan de in de ontvanger van de APC techniek invers opererende adaptieve voorspeller, die de synthetische spraak $\hat{s}(n)$ berekent uit vorige monsters $\hat{s}(n-1)$, $\hat{s}(n-2)$, ... en een gekwantiseerde versie van het zgn. residu signaal r(n) , zie fig. 5. Maar als dat zo is, dan is het residu signaal r(n) , zo men wil, ook een model voor het excitatie signaal in het spraakproduktie model. Fig. 9 toont 2 zulke korte residu signalen die typisch zijn voor respectievelijk een stemhebbend (fig. 9a) en een stemloos (fig. 9c) spraak segment.

Ter vergelijking zijn eveneens getoond de overeenkomstige korte vocoder excitatie signalen, nl. de puls trein (fig. 9b) en de witte ruis sekwentie (fig. 9d). De 2 typen excitatie signalen vertonen (uiteraard) zekere overeenkomsten, maar niettemin zijn de detail verschillen groot, vooral in stemhebbende segmenten. Bovendien zal de residu excitatie regelmatig kenmerken weerspiegelen van niet-stationaire overgangen, iets waarin de vocoder excitatie vorm nu te enen male niet duidelijk voorziet. Maar er is meer. De vocoder

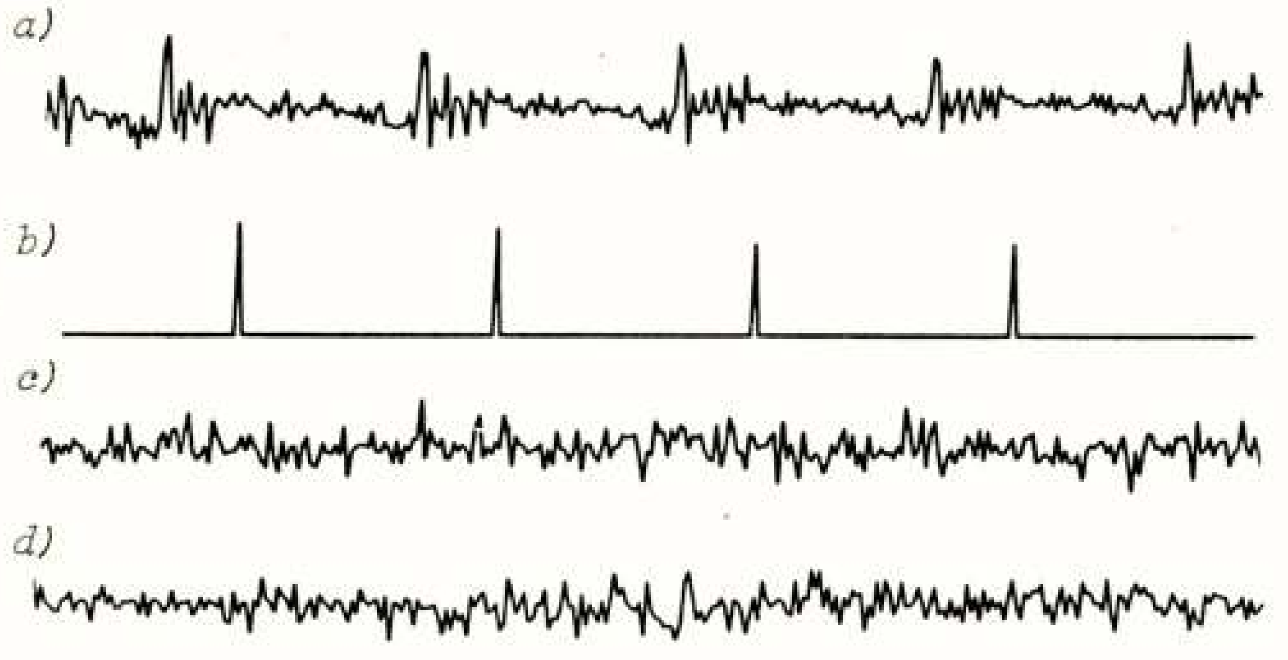


Fig. 9. Excitatie signalen: residu fragmenten (a) en overeenkomstige vocoder bronsignalen (b) en (d).

streeft ernaar de spraak via globale spectrale karakteriseringen te reconstrueren, zodat de golfvormen van de originele en de synthetische spraaksignalen geen gedetailleerde overeenkomst hoeven te vertonen. Met andere woorden, er is een hoorbaar gebrek aan coherentie. Daar staat tegenover, dat in de vocoder benadering alleen maar segment parameters gecodeerd worden, terwijl in bijv. de APC methode althans het residu signaal op monster basis gecodeerd wordt. Blok codering zou wel eens efficienter kunnen zijn. Kortom, door van de 2 soorten excitaties zowel de overeenkomsten als de verschillen tegen elkaar af te wegen moet het mogelijk zijn nieuwe excitatie signalen te construeren die zowel kwalitatief als kwantitatief intermediair zijn. Zo bijv. worden de pieken die duidelijk waarneembaar zijn in het APC residu signaal van stemhebbende spraaksegmenten, in de vocoder door een generator opgewekt. Ook in de adapatieve golfvormcoder kan men dit doen door aan de korte termijn segment voorspeller een periodiciteits- of pitch voorspeller toe te voegen. Daardoor verdwijnen de pieken min of meer uit het residu signaal dat dan meer het karakter heeft van een willekeurige sekwentie, zowel in stemhebbende als in stemloze segmenten. Bij de synthese worden de pieken dan weer opgewekt door ook deze voorspeller invers toe te passen bij de reconstructie van s(n) uit r(n) . Een andere betekenisvolle waarneming is deze. Het vocoder excitatie signaal is ofwel een witte ruis, ofwel een puls trein. Beide hebben een vlak spectrum, zie bijv. fig. 8b voor de puls trein. Ook het APC residu signaal heeft een vrij vlak (segment) spectrum. Dit feit kan worden uitgebuit door slechts een deel van dit vlak spectrum, zeg het 1e kwart, te behouden en dit dan voor reconstructie doeleinden (3 keer) te kopieren tot een volledig (vlak) spectrum.

Deze techniek is getoond in fig. 10 voor de zgn. residu basis band coderingsmethode (BBC). Een goed ontworpen BBC algoritme kan spraak comprimeren tot 8 a 10 kbit/s.

In zowel de vocoder als de BBC wordt het enig exacte excitatie signaal, nl. het (APC) residu signaal, vervangen door een substituut signaal dat op grond van spectrale eigenschappen geconstrueerd wordt. In beide

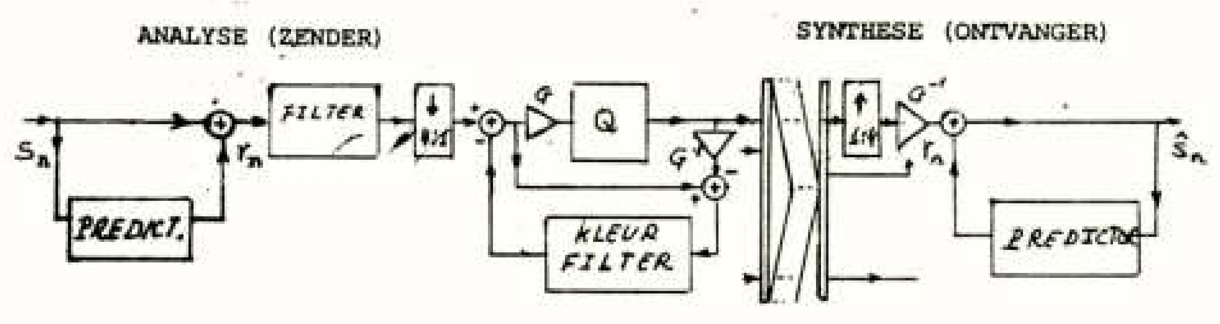


Fig. 10. Residu Base Band Codèr systeem.

gevallen is de golfvorm overeenkomst tussen de originele en de synthetische spraak niet gegarandeerd. Een voor de hand liggende vraag is dan deze: is het mogelijk een geschikt excitatie signaal te construeren zodat het synthetisch spraaksignaal zo min mogelijk afwijkt van het origineel signaal. Een dergelijke compressie methode zou er dan uitzien als geschetst in fig. 11. Daarin wordt een excitatie generator zodanig gestuurd dat het over een zeker tijdsinterval gemeten gewogen verschil e(n) tussen de natuurlijke en de synthetische spraak minimaal is in de zin der kleinste kwadraten.

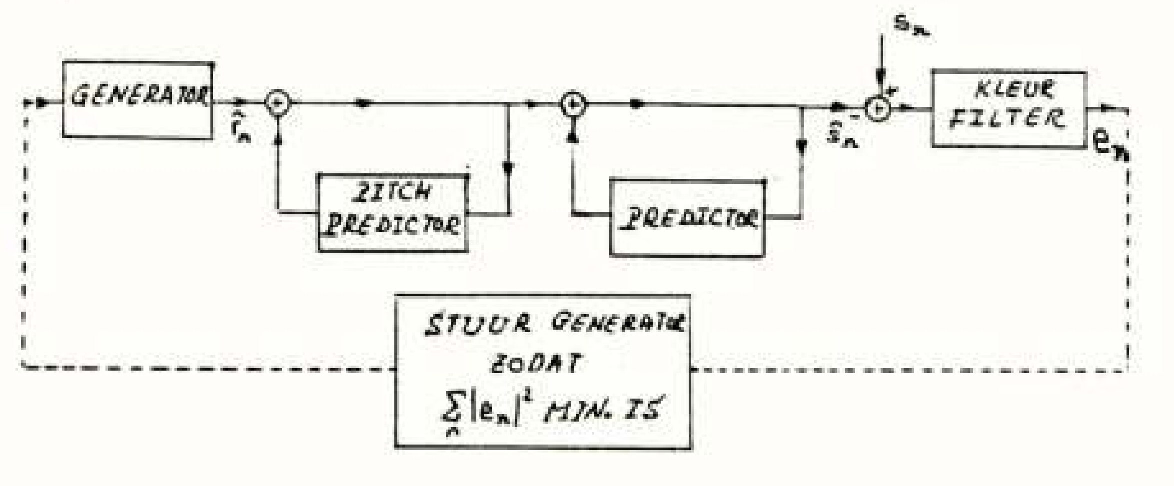


Fig. 11. Golfvorm codering met geconditioneerde excitatie generator.

Dit is uiteraard geen momentane procedure omdat 'gezocht' moet worden naar de optimale segment (blok) excitatie en het is dus zaak om de daarmee samenhangende vertraging binnen de perken te houden. Met deze methode zijn we beland bij de meest recente ontwikkelingen op het gebied van spraakcompressie. Afhankelijk van de manier waarop de aard van het vereiste excitatiesignaal geinterpreteerd wordt, kan het systeem volgens fig. 11 verschillende zoekprocedures opleveren. Binnen het kader van dit overzichtsverhaal kunnen we slechts zeer in het kort op een paar van deze ingaan.

Een van deze zienswijzen is een stochastische. Een spraaksignaal kan immers worden geinterpreteerd als een realisatie van een Gaussisch proces, met langzaam varierende spectrale eigenschappen. Zulk een proces kan worden gegenereerd, door een toevals proces met Gaussische kans-dichtheid te filteren met een langzaam varierend filter. Voor spraak is dit filter bekend, het bestaat uit de 2 voorspellers uit fig. 11. De resterende taak is dan het vinden van de beste reeks van toevalsamplituden die het gewogen verschilsignaal e(n) minimaliseert. Bij deze zgn multi-pad codering zijn veel gebruikte zoek strategieen de boom- en de tralie coders. In bijv. binaire boomcodering, zie fig. 12, dragen de takken van de boom een of meer toevals-amplituden en wordt het opklimmen en het afdalen aangegeven met een enkel symbool.

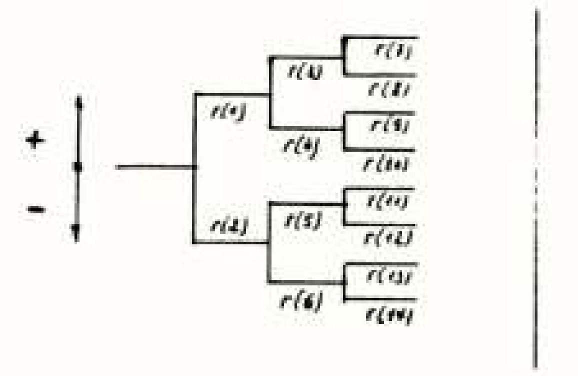


FIg. 12. Binaire boom met gaussische populatie.

Daarmee kan de ontvanger de bedoelde sekwentie in de eigen (identieke) boom terugvinden. Uitputtend uitkammen van de boom is een exponentieel karwei en is dus uitgesloten. Maar er bestaan wel sub-optimale lineaire zoekprocedures die het goed doen. Met zulke compressie methoden kunnen de excitatiesignalen met 1 of minder bits/monster gecodeerd worden. Daarmee worden bitfrekwenties rond de 8 kbit/s, en lager, bereikt.

Een tweede zienswijze is een deterministische. De redenering hier is, heel in het kort, dat de vocoder spraakkwaliteit onder de maat blijft omdat bijv. voor stemhebbende spraak de puls trein niet coherent is met het spraaksignaal, omdat ook te veel excitatie monsters een amplitude gelijk aan nul hebben, en omdat verder overgangsverschijnselen gebrekkig vertolkt worden door het excitatiesignaal. Toch kunnen vele van de kleinere amplituden in het APC residu signaal nul gemaakt worden. Dit zou bijvoorbeeld kunnen met een kwantiseringskarakteristiek als getoond in fig. 13, maar dit levert verklaarbaar geen goede resultaten op.

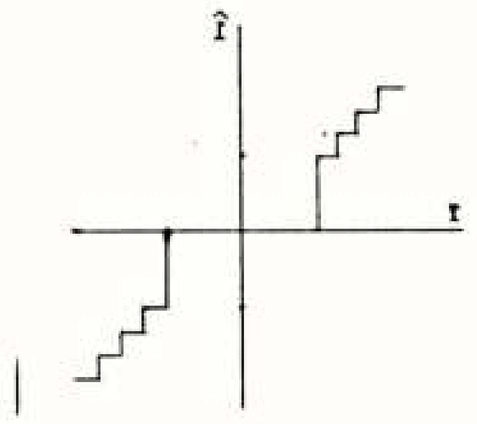


Fig. 13. Center-clipper kwantiseringskarakteristiek.

Er zit dus niets anders op dan te zoeken naar de plaats en de amplitude van een zo klein mogelijk aantal vermogen dragende monsters in een excitatie signaal dat verder alleen maar monsters met amplitude nul heeft en dat het gewogen verschil signaal e(n) minimaliseert. Ook hier is een optimale oplossing uitgesloten en wordt een sub-optimale zoekprocedure gevolgd die de betreffende monsters in elk minimalisatie geval een voor een opspoort. Met een 4-tal pulsen per 5 msec worden bitfrekwentie gehaald van 8 kbit/s, inclusief de filter parameter code, waarbij de synthetische spraak nauwelijks te onderscheiden is van de natuurlijke referentie spraak. Fig. 14 toont voor deze zgn. multi-puls excitatie methode (MPE), een synthetisch fragment (c) en de bijbehorende excitatie (b) en referentie (a) fragmenten.

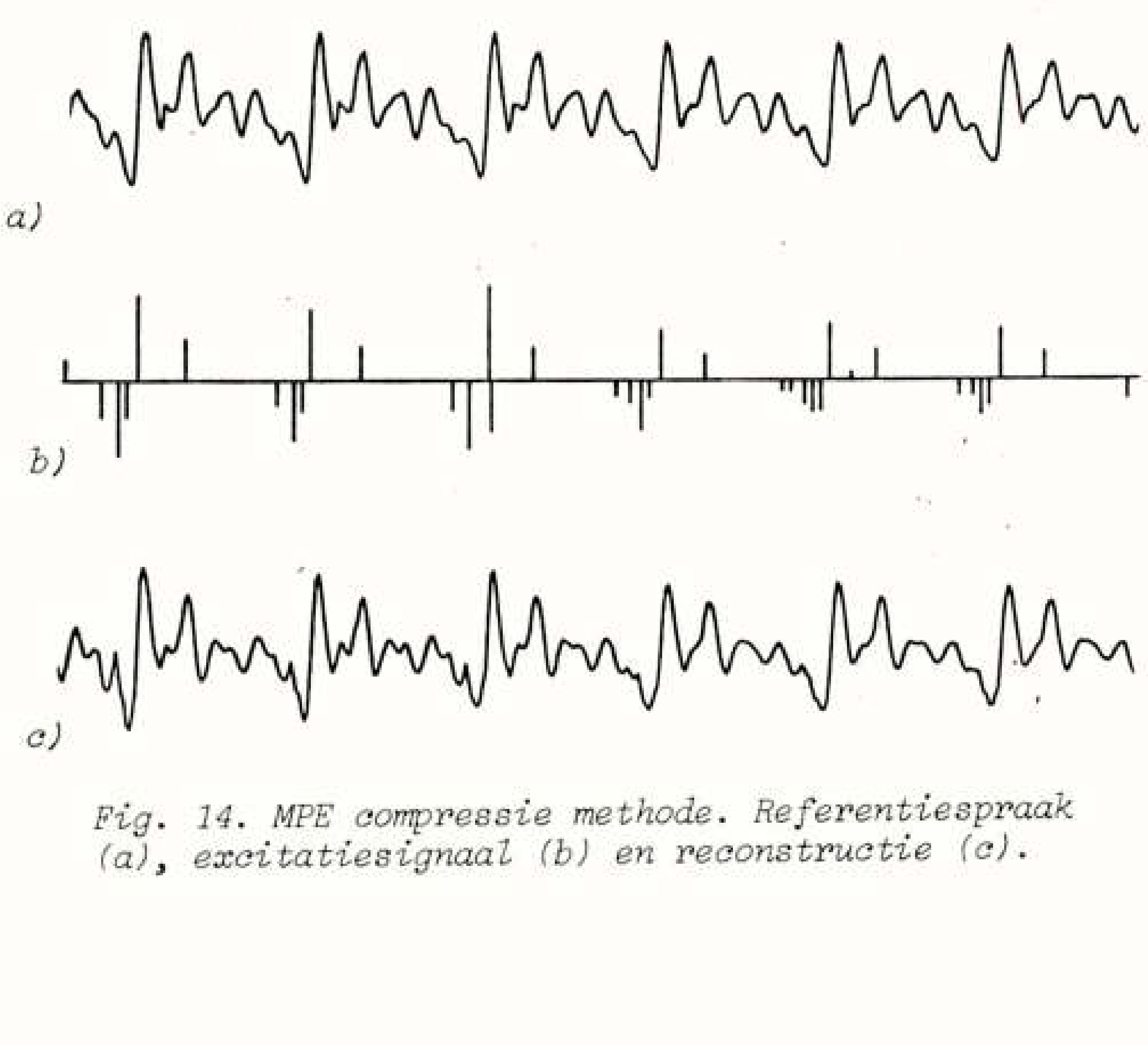


Fig. 14. MPE compressie methode. Referentiespraak (a), excitatiesignaal (b) en reconstructie (c).

Deze uiterst gespecialiseerde algoritmen zijn behoorlijk reken intensief. Maar het ligt in de lijn der verwachtingen, dat ze rekentechnisch geoptimaliseerd zullen gaan worden en dat, bovendien, in de nabije toekomst digitale signaal processor chips beschikbaar zullen komen waarin ook deze algoritmen in reele tijd geimplementeerd zullen kunnen worden.

VECTOR KWANTISATIE

In vrijwel alle genoemde coderingmethoden zijn compressie en reductie (d.i. kwantisatie) onafhankelijke stappen in de procedure, en is de kwantisatie scalair : zowel de residu (of de excitatie) monsters als de predictie parameters worden elk afzonderlijk gekwantiseerd. Dit is in het algemeen niet optimaal omdat de elementen in de te kwantiseren sequenties vrijwel altijd gecorreleerd zijn. Dit is overduidelijk het geval bij gewone PCM (96kbit/s bij 8kHz bemonstering) waarin de scalaire kwantisering geheel voorbij gaat aan de signaal redundantie.

Tegenover scalaire kwantisatie staat vector kwantisatie. De reeds genoemde (adaptieve) transformatie codering is in feite een vector kwantiseringmethode. Daarin worden de transformatie coefficienten weliswaar scalair gekwantiseerd, maar de blok (of vector) transformatie werkt decorrelerend,zodat de transformatie coefficienten (bij optimale transformatie) niet gecorreleerd, en dus optimaal scalair kwantiseerbaar zijn.

Vector kwantisering is niet wezenlijk anders dan scalaire kwantisering maar is implementatie technisch (in principe) veel complexer. In beide gevallen gaat het om de afbeelding van een element (scalar of vector) op een eindig alfabet van code woorden. Bij vector kwantisering heet deze eindige verzameling het codeboek (te vergelijken met de kwantisatie niveaus bij scalaire kwantisering). Het samenstellen van code boeken is op zich een probleem waar we hier niet op in kunnen gaan. Maar afgezien daarvan zijn de problemen ook niet gering. Want het benodige geheugen voor het opslaan van het code boek en de complexiteit van de zoekprocedure zijn exponentieel afhankelijk van de vectordimensie (voor een gegeven bit frekwentie). Practisch gesproken legt dit stringente beperkingen op, hetzij in termen van het boek volume (aantal en dimensie van de vectoren), hetzij in termen van de structuur van het codeboek, dit wil zeggen van de zoekstrategie. Zo is een grotere vector dimensie mogelijk wanneer sub-optimaal wordt gezocht .De kwaliteitswinst (langere vectoren) is echter maar beperkt (sub optimaliteit),de complexiteitsreductie kan evenwel aanzienlijk zijn, bijv. product-complexiteit (i.p.v exponentiele complexiteit). Een veel gebruikte structuur is de (niet noodzakelijk binaire) boom structuur.(Een andere mogelijkheid is het opsplitsen van het boek in een sequentie van deelboeken). Boom codering werd al genoemd bij de bespreking van de excitatie coderingsmethoden. Daar echter waren de boomtakken bevolkt met scalars, bij vector kwantisering zijn ze bevolkt met vectoren en geeft het getrokken pad doorheen de boom aan via welke weg de (sub optimale) vector (op een van de uiterste takken van de boom) bereikt kan worden. Het gevolgde pad is meteen het codewoord voor de vector.

Om een enkele indicatie te geven van de kracht van vector kwantisatie geven we ,tot slot, nog een vergelijking tussen een scalaire en een vectorkwantisatie van de kanaal parameters in een LPC vocoder. Voor een 'geringe fout' in de schatting van

het kanaalspectrum kost een scalaire kwantisering van deze parameters zo'n 27 bits per segment (10-20 ms) meer dan bij niet gestructureerde vector kwantisering en 20 bits per segment meer dan bij binaire boom vector kwantisering. Merk op dat de gestructureerde (en dus sub optimale) binaire zoek procedure slechts 7 bits/segment meer kost dan de optimale (niet gestructureerde) procedure. De complexiteiten echter (in termen van zoektijden bij computer simulaties) verhouden zich als minuten tot uren.

REFERENTIES

1. J.L. Flanagan et al., Speech Coding, IEEE Trans. on Communications, vol. COM-27, No.4, 710-737, 1979.

2. J.M. Tribolet en R.E. Crochiere, Frequency domain Coding of Speech, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No.5, 512-530, 1979.

3. J.D. Markel en A.H. Gray Jr., Linear Prediction of Speech, Springer Verlag, New York, 1976.

4. L.R. Rabiner en R.W. Schafer, Digital Processing of Speech signals, Prentice-Hall, New Jersey, 1978.

5. Special Issue on bit rate reduction and speech interpolation, IEEE Trans.on Communciations, COM-30, No. 4, April 1982.

6. A.Gersho en V.Cupermay, Vector Quantization: A Pattern-Matching Technique for Speech Coding , IEEE Communications Magazine, Vol.21,No.9,15-21,Dec.,1983.

Voordracht gehouden tijdens de 319e werkvergadering.

SPRAAKHERKENNING

Ir. L.J.P. van Heugten
DR. NEHER LABORATORIUM DER PTT

Speech recognition. A survey is given of Automatic Speech Recognition (ASR). By using vocal communication one can either "talk" information directly into the computer or control electromechanical systems by voice commands; it is the purpose of this paper to outline the most prominent problems in speech recognition and to describe frequently used methods and techniques to solve these problems. Attention is paid to isolated word and connected speech recognizers and to speech understanding systems.

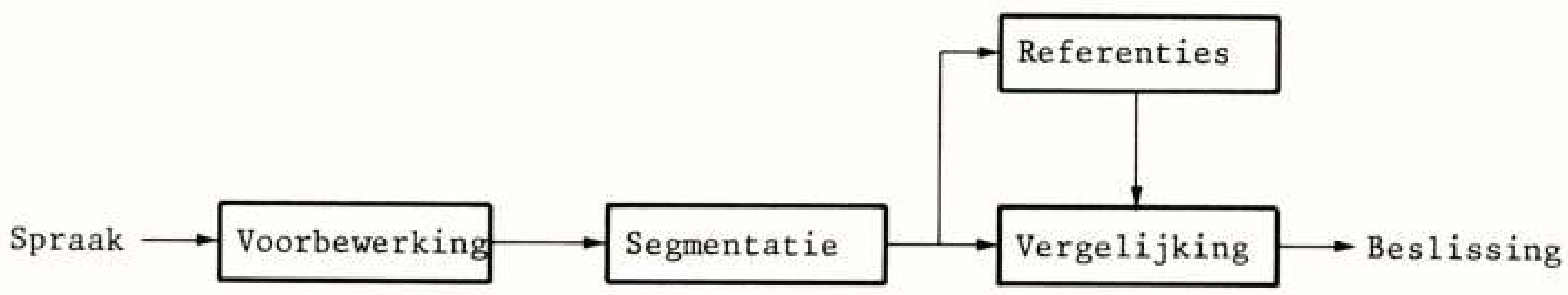
1. INLEIDING

Nu niet meer alleen in een technische omgeving, maar op steeds meer andere gebieden van de samenleving computers worden toegepast, wordt het steeds duidelijker dat de mens-machine communicatie niet aan de mens is aangepast.
Het is gelukkig niet meer zo dat er ponskaarten of -banden nodig zijn om programma's en dergelijke in te voeren, maar de interactie met toetsenbord en beeldscherm is niet erg mens-vriendelijk.
Het gangbare communicatiemiddel tussen mensen onderling is SPRAAK en als machines nu over een of andere vorm van luistervaardigheid zouden beschikken, zodat spraak als invoermedium gebruikt kan worden, zou een lang gekoesterde wens in vervulling gaan. Talloze toepassingen op velerlei gebied zijn denkbaar als spraak automatisch herkend wordt. Als het ook nog mogelijk wordt spraak via de telefoon in te voeren, zijn de toepassingen voor iedereen bereikbaar.
Al meer dan dertig jaar geleden werd de eerste "spraakherkenner" gemaakt: een speelgoedhondje genaamd "Radio Rex", dat op moest springen als zijn naam genoemd werd. (Het zal niemand verbazen, dat ook andere woorden het beestje konden laten springen en dat niet iedereen het voor elkaar kreeg).
De ontwikkeling van systemen die spraak herkennen, begrijpen en interpreteren is verder gegaan en momenteel worden op diverse plaatsen in de wereld al eenvoudige spraakherkenningssystemen toegepast. Er zal echter nog heel wat onderzoek verricht moeten worden, voordat er een machine ontwikkeld wordt, die ook maar een benadering is van de menselijke spraakherkenningscapaciteit. Voordat de "inspreekbare typemachine" een feit is, moeten er nog veel problemen worden opgelost.
Het immense probleem van de automatische spraakherkenning hoeft gelukkig niet in een keer in zijn geheel te worden opgelost. Door beperkingen op te leggen aan de gebruikers (de mens bezit een groot vermogen zich aan te passen aan de machine) kan het herkenningssysteem relatief simpel gehouden worden.
De problemen die optreden bij automatische spraakherkenning liggen in de aard van het spraaksignaal zelf. Een spraaksignaal is in principe eenmalig en het is de uiteindelijke uitkomst van een zeer complex proces, waarbij allerlei soorten informatie (bijvoorbeeld woordkennis, grammatica, syntaxis, uitspraakregels en intonatiepatronen) een rol spelen. Ook wordt het signaal beïnvloed door indirecte aspecten zoals de gemoedstoestand van de spreker, de fysieke bouw van de spraakorganen en dergelijke.
Het feit dat een spraaksignaal een eenmalige uiting is, impliceert dat een bepaald woord, diverse keren door een spreker uitgesproken, niet twee keer exact hetzelfde zal zijn. Uitspraakverschillen worden nog groter als een woord door verschillende sprekers wordt gezegd. Worden de woorden niet afzonderlijk uitgesproken, maar in een zin, dan zullen de begin- en eindklanken van het woord beïnvloed worden door de er omheen liggende klanken (co-articulatie). Herkenning wordt dus steeds moeilijker. Daar komt nog bij dat het in lopende spraak niet altijd mogelijk is aan te geven waar een woord begint en eindigt. Het is soms zelfs geheel onmogelijk woorden te segmenteren als men een op zichzelf staande zin hoort en geen extra informatie heeft over bijvoorbeeld de context of het gespreksonderwerp. Voorbeelden zijn woordparen die akoestisch weinig verschil vertonen, zoals "lief autootje" en "lief fotootje", en combinaties van woorden die "aan elkaar geplakt" worden en niet direct herkend worden zoals bijvoorbeeld "Loe

en sien ogen net".
Afhankelijk van de restricties die men aan de in te voeren spraak stelt, worden de systemen onderscheiden in losse woord-herkenners (isolated word recognizers), herkenners voor aaneengesproken worden (connected word recognizers) en lopende spraak-herkenners (continuous speech recognizers). Daarnaast kent men de spraakbegrijpende systemen (speech understanding systems), waarbij het niet de bedoeling is het spraaksignaal volledig te vertalen in een tekst, maar waar het gaat om de betekenis van de boodschap, de interpretatie hiervan om daarop actie te ondernemen. Alle drie de genoemde klassen kunnen in principe deel uitmaken van een spraakbegrijpend systeem; in het algemeen wordt de term gebruikt voor de grotere herkenningssystemen voor continue spraak, die een gespecificeerde taak uitvoeren
In dit artikel zullen de technieken, die bij spraakherkenning in gebruik zijn geschetst worden en zal een blik in de toekomst geworpen worden.

2. LOSSE WOORD-HERKENNING

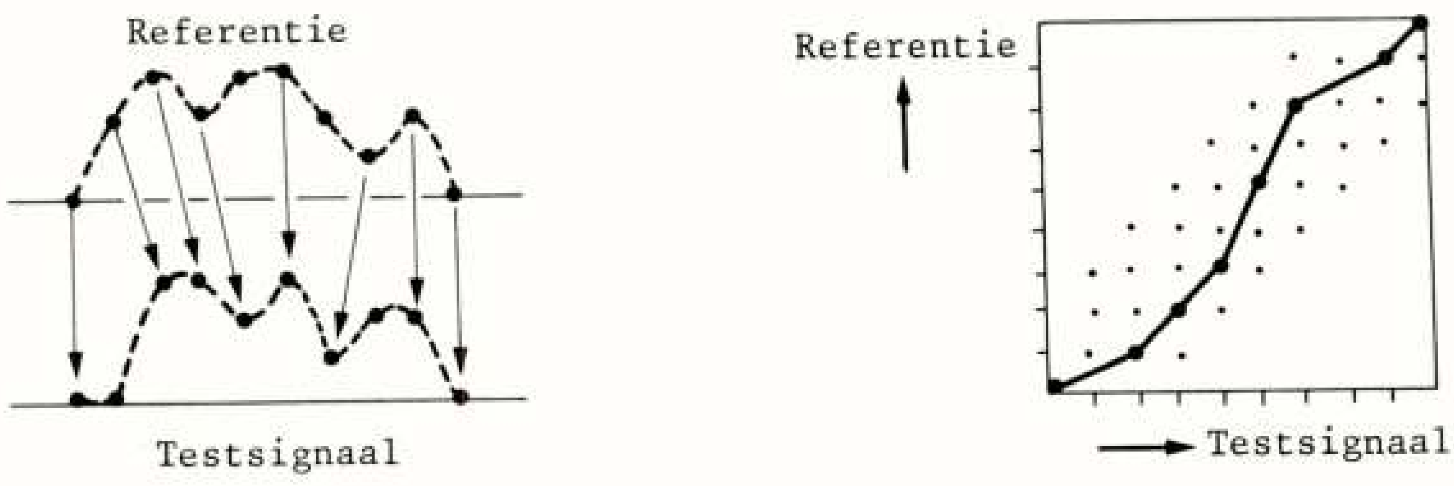
Alvorens men tot spraakherkenning kan overgaan, moet het akoestische signaal in een voor de computer geschikte representatie worden omgezet; een representatie waarbij de specifieke spraakkenmerken zomin mogelijk geheugenruimte beslaan.
Voorbewerkingstechnieken die worden toegepast kunnen zowel in het tijddomein (bijvoorbeeld het tellen van het aantal nuldoorgangen van het signaal) als in het frequentiedomein (bijvoorbeeld fourier-analyse) werken.
De twee meest toegepaste methoden voor de bewerking zijn de LPC-analyse en de filtering met behulp van bandfilters. Met beide methoden tracht men uit een stukje van het spraaksignaal (bijvoorbeeld van 20 ms) de spectrale informatie te extraheren. De spraak is na deze bewerking omgezet in een reeks kenmerkvectoren. (voor elk tijdinterval een set getallen), waarmee verder gewerkt wordt.
Na de voorbewerking vindt segmentatie plaats: hierbij wordt, met gebruikmaking van de stiltes tussen de woorden, de spraak gesplitst in de afzonderlijke woorden. De gebruikelijke term hiervoor is "end-point-detection". Daar er ook in het woord stiltes kunnen voorkomen (bijvoorbeeld bij het begin van een plofklank /k/, /t/ e.d.) mogen de stiltes niet te kort zijn. Het onbekende woordpatroon wordt dan vergeleken met de in de leerfase opgenomen referenties en er wordt beslist welk woord ingesproken werd.
De opbouw van een losse woord-herkenner is in figuur 1 schematisch weergegeven [ICASSP, 1983].



FIGUUR 1 : Blokschema losse woord-herkenner.

De referentie kan op verschillende manieren worden opgebouwd. De eenvoudigste methode is van elke voorkomend woord een uitspraak als referentie te nemen. Om rekening te houden met de variabiliteit in het spreken, zullen er meer uitspraken gebruikt worden om een referentie te bepalen via een middelings- of een clusteringsmethode. Als van een bepaald woord verschillende uitspraken bestaan (bijvoorbeeld "zeven" en "zeuven") is het zinvol meer dan één referentie per woord te bepalen. Ook als er meerdere personen van het systeem gebruik maken, zullen er verschillende referenties per woord zijn. Duidelijk is dat het aantal referenties niet onbeperkt kan toenemen: hoe groter de vocabulaire wordt, hoe meer geheugenruimte nodig is, hoe meer rekentijd er wordt gebruikt, hoe groter de kans is op onderlinge verwisselingen en hoe meer werk het is om de referenties van een nieuwe spreker te maken. Door de ontwikkelingen in de micro-elektronica tellen de eerste twee bezwaren steeds minder mee. Om aan het laatste bezwaar tegemoet te komen tracht men transformatieregels op te stellen om het systeem aan een nieuwe spreker aan te passen (sprekeradaptie).
Om een zo goed mogelijke herkenning te krijgen, zullen optredende uitspraakverschillen zo mogelijk moeten worden teniet gedaan. Een zeer populaire methode is de "dynamic-time-warping" [Sakoe, 1978]. Deze methode is in staat in de berekening van het verschil tussen de test- en de referentie-uitspraak te corrigeren voor de tempovariaties in de spraak, die in het algemeen niet-lineair blijken te zijn, doordat ze alleen kunnen optreden bij bepaalde klanken (bijvoorbeeld klinkers kunnen wel ingekort worden, maar plofklanken niet). De test- en referentie-uitspraak worden zo mogelijk op elkaar afgebeeld door tijdstippen over te slaan of te

introduceren ("matching") en wordt de hierbij behorende minimale afstand berekend. Het principe is geschetst in figuur 2.



FIGUUR 2 : Principe van dynamic time warping (DTW).

Als alle benodigde technieken worden aangewend zal een herkenningssysteem dat uitsluitend gebruik maakt van de akoestische eigenschappen van spraak, nooit honderd procent accuraat kunnen werken. Onvolkomenheden in de woordherkenning (bijvoorbeeld veroorzaakt door een slordige uitspraak) kunnen worden verbeterd door voor de in te spreken woorden een syntaxis (grammaticale structuur) te definiëren. Een voorbeeld is een inspreekbare rekenmachine waarbij alleen de structuur

getal--*bewerking*--*getal*--*uitkomst*

voorkomt. Als het eerste getal is ingevoerd kan alleen een bewerking volgen; er hoeft hier dan alleen uitgemaakt te worden of het volgende woord gelijk is aan "plus", "min", "maal" of "gedeeld-door". In feite wordt door de structuur de woordkeuze op een bepaald punt in de "zin" beperkt tot een deelset.

Het spreken is niet erg natuurlijk en kan op den duur erg vervelend worden. De eerste stap om de herkenners wat mens-vriendelijker te maken is toe te staan de woorden zonder pauzes op te noemen. Directe toepassing is mogelijk bij cijferreeksen (telefoon-, rekeningnummers of bedragen). Bij de bestaande systemen voor herkenning van aaneengesproken woorden gaat men uit van de referenties van de losse woordherkenning, met aanpassingen voor de segmentatie en de "matching"-methoden, eventueel nog aangevuld met regels om voor woordoverlapping te corrigeren.

Momenteel zijn er voor sprekerafhankelijke toepassingen commerciële losse woord-herkenners te koop, in prijs variërend van $500 tot $10.000, met herkenningspercentages van resp. 92% tot 98% (percentages opgegeven door de fabrikant). Een sprekeronafhankelijke herkenner voor 10 (aaneengesproken woorden) tot 40 (losse) woorden kost $80.000 [Lea, 1983].

De prijsverschillen worden onder meer beïnvloed door de robuustheid van de verschillende systemen in "moeilijke" omstandigheden (die in dit artikel buiten beschouwing zijn gelaten), zoals een lawaaiige omgeving, een smalle bandbreedte, verbindingen via de telefoon en extra mogelijkheden zoals "real-time" verwerking, muliple-access (diverse personen tegelijk), gebruiksgemak en dergelijke. In de praktijk is een aantal systemen in gebruik voor het vervoer van bagage op luchthavens, voor de kwaliteitscontrole in de industrie, voor het sorteren van pakketpost en invoeren van gegevens van luchtfoto's.

3. SPRAAKHERKENNENDE EN SPRAAKBEGRIJPENDE SYSTEMEN

De ontwikkeling van de sprekeronafhankelijke spraakherkenner voor een onbeperkte woordenschat kan niet gebaseerd worden op de losse woord-herkenning: de referentieset zal te groot worden omdat er eindeloos veel woordcombinaties mogelijk zijn met de hierbij optredende co-articulaties. Deze moeilijkheid kan worden omzeild door geen hele woorden, maar kleinere eenheden te herkennen.

In de fonetiek (het deel van de taalwetenschap dat zich bezighoudt met de bestudering van spraakklanken) bestaat het begrip foneem: de kleinste klankeenheid die woordonderscheidend werkt [Ladefoged, 1975]. Zoals in deze defintie naar voren komt is het begrip foneem gekoppeld aan taalkundige aspecten en niet aan akoestische. Een foneem kent dan ook door de invloed van omringende klanken verschillende realisaties (bijvoorbeeld de /l/ in "melk" en in "alaaf"); dit worden allofonen genoemd. In de Nederlandse taal onderscheidt men circa veertig fonemen en ongeveer honderd allofonen. Dit is een overzichtelijk aantal; aan de foneemherkenning zitten echter nog wat haken en ogen. Ten eerste is het karakter van de fonemen (tussen de klassen onderling) zeer verschillend (vergelijk bijvoorbeeld een klinker met een plofklank) zodat verschillende strategieën nodig zijn; de fonemen binnen een klasse zijn daarentegen zeer moeilijk te onderscheiden.

Ten tweede, en dit is veel belangrijker, spreken we niet in losse fonemen: er is tijd nodig om de stand van de tong, mond en de lippen te veranderen zodat we bij het spreken verglijden van het ene foneem in het andere, waarbij vaak nog voordat een foneem echt gevormd is, al wordt begonnen de spraakorganen te veranderen voor het volgende foneem. Proeven hebben uitgewezen dat de mens ook uit de omgeving van de klank veel informatie haalt: klinkers los uitgesproken (ze bevatten dan dus de overgangen naar stilte) worden zonder problemen herkend; als klinkers uit een zin worden "geknipt" en men laat ze in isolatie horen, wordt door proefpersonen slechts vijftig procent juist herkend!

Nemen we als herkenningseenheid niet de fonemen, maar juist de overgangen van het ene foneem naar het andere (difonen), dan wordt de informatie die hierin aanwezig is benut. Het aantal difonemen is ongeveer 1200. Gaan we nog een stapje verder dan komen we bij de halve of hele lettergrepen (demi-syllabe/syllabe). Hiervan zijn er omstreeks 2000. De segmentatie van spraak in halve lettergrepen kan in het ideale geval automatisch geschieden door te zoeken naar energietoppen en -dalen. Welke herkenningseenheid we ook gebruiken, er kan niet volstaan worden met één referentie per foneem, difoon of demi-syllabe, maar er zullen procedures nodig zijn voor de herkenning. Het is nog niet bekend wat de beste herkenningseenheid is, of welke combinatie de beste resultaten geeft. Een nadeel dat ze allemaal hebben is de sprekerafhankelijkheid. Aan dit probleem is nog weinig aandacht besteed. Er zijn wel enkele methoden ontwikkeld voor sprekernormalisatie maar hoe nu precies de sprekersinvloed uit het spraaksignaal te verwijderen is, is nog onbekend.

De uitkomst van de herkenner bestaat over het algemeen uit een of andere code, die het spraaksignaal representeert. Deze code moet verder behandeld worden om uiteindelijk te komen tot een tekst of een interpretatie. Zo komen we dus bij de spraakbegrijpende systemen. Eind jaren '60, begin jaren '70 waren er reeds enkele grote projecten aan de gang. Op diverse wijze trachtte men de taalkundige informatie te benutten. Zo introduceerde men Markov-modellen voor de dialoogbeschrijving en maakte men statistieken voor opeenvolging van klanken met bijbehorende kansen.

Geweldig ambiteus werd in 1971 door het Advanced Research Project Agency (ARPA) van het Amerikaanse ministerie van defensie het Speech Understanding Research (SUR)-project gestart, met als doel een doorbraak te forceren in de spraakverwerking. De studiegroep benadrukte het concept van spraakbegrijpende ten opzichte van spraakherkennende systemen en definieerde als doel het ontwerpen van een systeem dat moest voldoen aan [Klatt, 1977]:

"Accept continuous speech, from many cooperative speakers, in a quiet room, with a good microphone, with slight adjustments for each speaker, accepting 1000 words, using an artifical syntax, yielding less than 10% semantic error, in a few times real time".

Vijf jaar en vijftien miljoen dollar later werden vier systemen gedemonstreerd, waarvan er één volledig aan de specificaties voldeed (de "HARPY" [W.A. Lea, 1983]). Naast dit geweldige resultaat is er tijdens het project op diverse gebieden ervaring opgedaan en belangrijke vooruitgang geboekt. Voorbeelden hiervan zijn: systeemorganisatie, grammatica-ontwerp, controlestrategieën, semantiek en context, fonetische analyse en gebruik van statistische kenmerken.

Soms wordt weleens gedacht, dat het onderzoek van de spraakherkenning alleen bestond uit het ARPA-SUR project, maar vrijwel alle grote laboratoria zijn bezig met onderzoek op het gebied van de spraakherkenning; in de Verenigde Staten zijn dat bijvoorbeeld Bell labs, IBM, TI e.a., in Japan o.a. NEC, NTT, Matsushito, en verder laboratoria in Europa, Rusland enzovoort.

Momenteel worden enkele spraakbegrijpende systemen in de praktijk beproefd o.a. het luchtvaart-informatiesysteem van Bell labs in Amerika [Levinson, 1980] en het treinreserveringssysteem in Japan [Nakatsu, 1978 en Sakoe, 1978].

4. TOEKOMSTVERWACHTINGEN

Steeds vaker komt men in (vak-)tijdschriften artikelen en advertenties tegen van grote en kleine systemen, die spraak als in- en/of uitvoergrootheid gebruiken. Dit betreft dan voornamelijk losse woord-herkenners en synthese-chips, die door het succes in de micro-elektronica op grote schaal voor weinig geld geproduceerd kunnen worden. Op korte termijn zullen deze eenvoudige spraaksystemen in allerlei toepassingen op de consumentenmarkt (spelletjes, speelgoed, elektrische apparaten en in de autobranche) gebruikt worden. Het is echter te betwijfelen of het publiek behoefte heeft aan het gebruik van spraak in allerlei snufjes, waar de spraak niet echt als medium benut wordt.

Op langere termijn is voor de lopende spraakherkenning een rol weggelegd in de kantoorautomatisering, in informatiesystemen via de telefoon, spraakbestuurde apparatuur (gehandicapten) en dergelijke. Verwacht wordt dat sprekeronafhankelijkheid over vijf jaar werkelijkheid is geworden en dat in de jaren '90 in de kantooromgeving de spraakverwerkingsapparatuur even normaal zal zijn als de typemachine nu. De vraag of het ooit mogelijk zal zijn de perfecte spraakherkenner te bouwen blijft onbeantwoord. Zeker is dat er in de huidige kennis nog grote hiaten zitten en dat de technologie momenteel ontoereikend is. Het toekomstig onderzoek

op het gebied van spraakherkenning mag zich dan ook niet beperken tot het op grote schaal toepassen van de bestaande herkenners, vervolmaken van huidige systemen en technieken en het ontwikkelen van systemen gebaseerd op bekende modellen. Zowel fundamenteel als toegepast onderzoek op het gebied van taalkunde en taaltechnologie moeten de noodzakelijke inzichten en nieuwe technieken verschaffen om stap voor stap "DE HERKENNER" te ontwikkelen.

5. REFERENTIES

Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP 1983, Boston.

D.H. Klatt, 1977.
"A review of the ARPA speech understanding systems", J. Acoust. Soc. America, Vol. 62, No. 6, pp 1345-1366.

P. Ladefoged, 1975.
"A course in phonetics", New York, Harcourt Brace Jovanovich.

W.A. Lea, 1983
"Selecting the best speech recognizers for the job", Speech Technology, Vol. 1, No.4, pp 10-29.

S.E. Levinson, K.L. Shipley, 1980.
"A conversational-mode airline information- and reservation system using speech input and output", Bell System Techn. J., Vol. 59, No. 1, pp 119-137.

B. Lowerre, 1977.
"The Harpy speech recognition systems", Ph.D. dissertation, Computer Science Dept., Carnegie-Mellon U.

R. Nakatsu, M. Kohda, 1978.
"An acoustic processor in a conversational speech recognition system", Review of the electrical communication laboratories, Vol. 26, No. 11-12, pp 1486-1504.

H. Sakoe and S. Chiba, 1978
"Dynamic programming alorithm optimization for spoken word recognition", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-26, No. 1, pp 43-49.

K. Shikano, M. Kohda, 1978.
"A linguistic processor in a conversational speech recognition system", Review of the electrical communication laboratories, Vol. 26, No. 11-12, pp 1505-1520.

Voordracht gehouden tijdens de 319e werkvergadering.

FONEMATISERING VAN GESCHREVEN TAAL

M. Boot

RU Utrecht
Afdeling Computerlinguistiek van de vakgroep TTCL
Wilhelminapark 11
3581 NC Utrecht

CONVERSION OF WRITTEN TEXT INTO PHONEMES. In order to construct automatic reading machines, we first have to convert the written text into phonemes. This paper discusses the problems encountered during the design of such an algorithm for the Dutch language.

Het hart van iedere voorleesmachine is het computer programma voor de omzetting van de letters in fonemen. Er zijn wat Nederlands betreft 2 generatie omzetters. Ik zal die in het onderstaande nader bespreken. Vooraf echter een korte aanduiding van de problematiek die men ontmoet bij het fonematiseren via de computer.

De spelling van het Nederlands ligt redelijk dicht bij de uitspraak. Toch zijn er aanmerkelijke problemen. Men denke bijvoorbeeld aan de letter E in de volgende woorden:

E=ə	E=EE	E=E
geval	gevel	
gezel	gesel	
gespot		gespen
bestelen		beste
beneveld	benen	
vergezellen		vergezicht
verrekenen		verrekijker
versmaden		versmaat
verflauwen		verflaag
verzenden		verzen
verversen		verven
tegelijk	tegelbakker	
meteen	meten	metselen

Het spreekt voor zich dat een computer programma dat letters in fonemen moet omzetten in staat moet zijn om gevallen zoals boven gegeven te transcriberen. Het gaat hierbij om gewone, veel voorkomende Nederlandse woorden.

Eerste generatie omzetters

De eerste generatie omzetters werd ontwikkeld op het Mathematisch Instituut in Amsterdam. Dat gebeurde zo vanaf het midden van de jaren 60. Het ontwerp van de eerste generatie zag er als volgt uit:

1. Het hart van het ontwerp en tevens de eerste stap van het programma is een programma voor woordafbreking.

2. Nadat de opsplitsing in lettergrepen is gedaan wordt een serie woorddelen afgesplitst.

3. Het programma maakt gebruik van de klemtoon van de Nederlandse lettergreep.

4. Als de stappen 1 tot en met 3 zijn doorlopen vindt de omschrijving van letters naar fonemen plaats. Hierbij worden ook assimilatieregels toegepast.

Het meest opvallende aan dit ontwerp is dat de eigenlijke konversie afhankelijk is van 2 andere programma's, een voor woordafbreking en een voor klemtonen [5,6]. Die programma's zelf echter zijn allerminst op hun taak berekend. Men kan dat afleiden uit de afbreekfouten die dagelijks in de krant te vinden zijn. De dagbladpers gebruikt namelijk hetzelfde afbreekprogramma als de fonematiseerder van de eerste generatie. Het aantal foutieve afbrekingen beloopt daar tot 60% van de afgebroken woorden. Het klemtoonprogramma levert een nog lager korrekt percentage.

Resulaten van de eerste generatie

Er is een tweetal testen gepubliceerd. De eerste test bestond uit een toepassing van het programma op het Van Berckel corpus [2]. Alleen woorden met een frekwentie hoger dan 1 werden voor die test gebruikt. Van deze woorden waren er 219 verkeerd getranscribeerd. Volgens Kok [5] was dat 6%. De fouten worden als volgt verklaard:

1. 9 % wordt veroorzaakt door woordafbreking.

2. 16 % wordt verklaard doordat het geen Nederlandse woorden zijn.

3. 24 % wordt veroorzaakt doordat het afkortingen zijn en door fouten in het klemtoonprogramma.

4. 51 % kan niet worden verklaard en wordt als onoplosbaar gekenschetst [5].

In de laatste kategorie vallen woorden als:

tafel / tegenover / kabel / mevrouw / hoogleraar.

Het gaat hierbij om veel voorkomende woorden. Het probleem van de E lijkt niet opgelost.

De tweede test leverde de volgende resultaten op. Het ging om 11.280 woorden waarvan er 410 fout werden getranscribeerd. Bij de beoordeling werden andere kriteria aangelegd maar uit de literatuur wordt niet duidelijk welke dat waren. De foutenkategorisering gaf het volgende beeld.

1. 30 % wordt verklaard door het woordafbreek programma.

2. 20 % wordt verklaard doordat het geen Nederlandse woorden zijn.

3. 6 % wordt verklaard door weglating van letters.

4. 44 % kan niet worden verklaard.

In de laatste kategorie komen woorden voor van de volgende soort:

lichaam / herken / generaal / losliggend / bedelaar.

Op het eerste gezicht is men geneigd de goede score van 6 % een goed resultaat te noemen. Kijkt men echter nauwkeuriger dan blijkt dat de gehele tabel die ik heb voorgesteld als kwaliteitsmaat onder de onoplosbare gevallen te vallen. Ofwel met de eerste generatie klankomzetters is het probleem van de E niet op te lossen. De 6 % fouten is dus gezichtsbedrog.

Tweede generatie omzetters

Aan de eerste generatie ontbrak een grondige taalkundige analyse van het probleem. Verder was de eerste generatie gebaseerd op computer programma's die hun eigen taak niet korrekt konden oplossen. Deze lessen uit de eerste generàtie hebben geleid tot het ontwerp van de tweede generatie. In het kort ziet dat ontwerp er als volgt uit:

1. Baseer de regels in het computer programma op fonologische inzichten en niet op lettergreepsplitsers.

2. Baseer de implementatie op patroonherkenning en niet op andere programma's.

Fonologie

Wil men een computerprogramma maken voor fonematisering dan zal men de regelmaat moeten zien te ondekken die in de omzetting van tekst naar spraak zit. Die regelmaat vindt men beschreven in de fonologie. Deze tak van de taalkunde bepaalt welke fonemen tot een bepaalde taal behoren en welke betrekkingen tussen de fonemen zijn toegestaan. Veel van de regels die werden gevonden in de fonologie zijn speciaal van belang voor fonematiseerders omdat daarin de omgeving wordt beschreven die een specifiek foneem beinvloedt. Men denke aan assimilatie regels bij voorbeeld, op grond waarvan men de woorden LIEFDE en LIEFSTE achtereenvolgens uitspreekt als LI.VDə en LI.FSTə ([1], regels 4 en 21). Jammer genoeg beschrijft de fonologie niet alle regels die nodig zijn om een fonematiseerder te maken. Een voorbeeld hiervan is de regels voor de morfemen -IG en -LIJK die worden geinterpreteerd als -əG en LəK, waarop uitzonderingen bestaan als VERONGELIJKT. Dit soort regels vormt de basis van de tweede generatie klankomzetters die werd ontwikkelde in de afdeling computerlinguistiek van de vakgroep TTCL in Utrecht. Daarbij is duidelijk dat deze tweede generatie een kontekstgevoelig regelsysteem bevat. Anders gezegd: deze tweede generatie is gebaseerd op de herkenning van patronen.

Morfologie

Naast de kennis van fonologische regels moet het programma ook kennis hebben van de morfologie van de taal die moet worden herschreven. De klankomzetter moet suffixen en prefixen kunnen herkennen. Men denke aan de volgende morfemen: -TIE (-TSI. of SI.); -VER (VəR) -AIR (Σ:R). Een woordafbreekprogramma zou fouten morfemen aanwijzen. Men denke bij voorbeeld aan een woord als PAARDEN dat wordt afgebroken als PAAR- DEN. Verder denkend kan men zich voorstellen wat een fouten een dergelijk programma kan maken.

Fonologische kennis in het programma FONGRAF, het computer programma van de tweede generatie.

In het programma is die kennis opgeslagen die niet al aanwezig is in de Nederlandse orthografie. Zo is er een regel dat een woord in het Nederlands niet kan beginnen met NG. Die regel hoeft niet te worden geimplementeerd simpel omdat er geen woorden voorkomen die beginnen met die lettercombinatie. Zo ook de regel dat een woord in het Nederlands niet mag eindigen op een V. De volgende 8 fonologische regels bleken voldoende:

1: b, d, g, ----→ p, t, x / --#

voorbeeld: heb, hond, vlag

2: C C ----→ C

2 identieke konsonanten worden 1 konsonant

voorbeeld: katten

3: ŋ / a ' e ' i ' u ' o -----+

voorbeeld: bang, breng, slungel

4: Toevoeging van een j in woorden als

zeeen , buiig

0 ----→ j / ..ee ie' ui -----//

5: i ----→ j / ..aa! oo! oe ../

voorbeeld: maaien, dooier

6: p t k s f x ----→ b d q z v g .. . V -- b'd
a

een stemloze konsonant wordt stemhebbend

als B of D volgt.

voorbeeld: opdoen, uitdoen , zakdoek, huisdeur

6: b d g v z ----→ p t x f s / .. V --
b
C(C not B'D)

Ieder kombinatie van 2 konsonanten levert

2 stemloze konsonanten op.

voorbeeld: klubkas, badgast, hebzucht

7: t ----→ 0 / . $V\begin{Bmatrix}x\\s\end{Bmatrix}$ ----- j

T tussen consonant en j wordt weggelaten

voorbeeld: kistje, grachtje

8: t d ----→ 0 / ... n .. C ..
/ ... C ... C ...

voorbeeld: onverwachts, kinds, vriendschap

Met deze 8 regels is de harde kern gegeven van de klankomzetter van de tweede generatie. Men ziet de kennis is heel erg toegesneden. Ofwel: men heeft lang niet alle taalkundige kennis nodig die er is.

Behalve de kennis die uit de fonologie en de morfologie te halen is heeft de expert voor fonematisering echter ook de kennis nodig over de volgorde waarin de regels moeten worden toegepast. Zo zal regel 8 moeten worden uitgevoerd voor regel 6 en regel 6 weer voor regel 2. Een uitvoerige beschrijving van de regels is te vinden in [4].

Resultaten Tweede Generatie

In eerste instantie werden 6 testen uitgevoerd toen de expert voor klankomzetting eenmaal klaar was (FONGRAF). De testen bestonden alle uit klankomzettingen door FONGRAF. De eerste test werd gedaan op de 1000 meest frekwente woorden van het Uit Den Boogaart korpus [3]. De tweede werd uitgevoerd op 1000 woorden met meer dan een lettergreep uit het Amsterdamse korpus [2]. De derde test geschiedde op een korpus dat was gekozen uit de letter A van Koenen [7]. Alle woorden werden genomen, ook de niet Nederlandse woorden en infrekwentie woorden ; daarnaast bevatte dit korpus ook nog alle 'onoplosbare' gevallen uit de eerste generatie. De vierde test bestond uit een test op de letter E. De vijfde test werd uitgevoerd op een korpus dat was samengesteld bij de ontwikkeling van de FONGRAF zelf. De zesde test werd uitgevoerd op een random gekozen kranteartikel. De resultaten waren als volgt:

Test	aantal woorden	fouten	%	gem. woordlengte
1	1071	0	0.00	5.77
2	988	25	2.50	8.02
3	556	75	13.49	8.18
4	303	15	4.95	5.91
5	279	20	7.17	8.28
6	364	6	1.65	5.15

Ofwel de 1000 meest frekwente woorden van het Nederlands worden juist getranscribeerd. Het woordenboek korpus bevat de meeste fouten, wat te verwachten was. Een willekeurige tekst uit een krant bevat minder dan 2% fouten.

Sinds 1980 is FONGRAF verbeterd. De testen op de resultaten zijn in volle gang. Tot nu toe hebben ze het volgende opgeleverd:

Test	aantal woorden	% oude versie	% nieuwe versie
1	1252	6.70	1.68
2	306	4.24	0.32
3	927	7.00	1.30

Waaruit de konklusie kan worden getrokken dat FONGRAF verbeterbaar is en dat er nog geen onoplosbare gevallen zijn opgedoken.

REFERENTIES

1. B. van den Berg, Foniek van het Nederlands, Amsterdam, 1958.

2. Van Berckel e.a., Formal Properties of Newspaper Dutch, Amsterdam, 1963.

3. P.C. Uit den Boogaart, Woordfrekwenties in geschreven en gesproken Nederlands, Utrecht, 1975.

4. M.Boot, E.Maat, J.Renkers, A Model for Automated Phonemization, Utrecht, 1980.

5. G.H.A. Kok, Het automatisch omzetten van geschreven Nederlands in een fonetische notatie, Amsterdam, 1972.

6. Papp, Zsepe, Papers in Computational Linguistics, Budapest, 1971.

7. Koenen, Woordenboek der Nederlandse Taal, 1974.

Voordracht gehouden tijdens de 319e werkvergadering.

SPRAAKSYNTHESE: STAND VAN ZAKEN EN TOEKOMST

Ir. L.F. Willems
Instituut voor Perceptie Onderzoek

The state of the art of speech synthesis is described in this paper. The application of speech synthesis in speaking machines is coming nearer through the availability of speech synthesis chips. The text to speech conversion problem is, however, not yet solved satisfactorily.

1. INLEIDING

Spraak is voor ons mensen een heel natuurlijk communicatiemiddel. Wij maken er veelvuldig gebruik van en het is ook te verwachten, dat bij geavanceerde en mens-vriendelijke communicatie tussen mens en machine spraak een grote rol zal spelen. Spraaksynthese, het opwekken van kunstmatige spraakklanken, heeft hier het doel om boodschappen vanuit een apparaat te produceren, zodat de menselijke gebruiker ze kan verstaan en erop kan reageren.

De mens heeft al sinds lang de spraak bestudeerd en ook geprobeerd spraakklanken na te bootsen. Een van de eersten was Wolfgang von Kempelen, die in 1791 een spreekmachine construeerde, waarmee hij, zoals hij schreef:

'...alle Latijnse, Franse en Italiaanse woorden zonder uitzondering kon namaken...zoals bijv. Papa, Maman, Marianna, Maladie, enz. ...'

Von Kempelen had voor de bediening van zijn mechanische spreekmachine beide handen en de nodige vingervlugheid nodig om dit te kunnen presteren. Na de uitvinding van de telefoon en toen deze ingevoerd raakte ontstond belangstelling voor het (electrische) spraaksignaal van de kant van de telefooningenieurs.

In de dertiger jaren heeft Homer Dudley van de Bell Labs pionierswerk verricht. Hij maakte de Vocoder en de Voder. Rond die tijd werd de geluidsspectrograaf ontwikkeld, waarmee het veranderende spectrum als functie van de tijd zichtbaar kon worden gemaakt.

Na de tweede wereldoorlog was er op vele gebieden van de wetenschap een opleving, óók op het gebied van het spraakonderzoek. Er was toen grote belangstelling voor spraakproduktie (articulatie, akoestiek van het mondkanaal, synthetische spraak) en ook voor de waarneming van spraak door de mens (auditieve filtering, Motor Theory of Speech Perception, enz.). In het begin van de 70-er jaren is de LPC-techniek ontwikkeld en nu beleven we de tijd van de stormachtige ontwikkeling van de electronica, waardoor fantastische mogelijkheden beschikbaar komen (computing power, (V)LSI-schakelingen, etc.).

We zullen in dit artikel allereerst ingaan op een aantal doorsnijdingen die men kan maken in het gebied van de spraaksynthese. De eerste doorsnijding heeft te maken met de techniek:

- golfvormcodering, versus:
- resynthese van geanalyseerde spraak, versus:
- spraaksynthese door regels.

Een tweede doorsnijding heeft te maken met toepassingsgebieden:

- vaste boodschappen
- variabele boodschappen
- willekeurige tekst uitspreken

Een derde doorsnijding is:

- complexiteit
- benodigde bitrate of geheugencapaciteit
- spraakkwaliteit

Vervolgens willen we nagaan hoe spraakklanken door de mens gemaakt worden om daaruit mogelijkerwijs inspiratie op te doen voor de manier waarop we spraakklanken kunnen nabootsen.

Het zwaartepunt zal vervolgens liggen bij de middelen om spraak te resynthetiseren en de mogelijkheid voor spraaksynthese door regels.

2. ENKELE ALGEMENE OPMERKINGEN OVER SPRAAKSYNTHESE

Voordat we zijn ingegaan op de methoden om spraak te synthetiseren willen we enkele algemene opmerkingen maken die de verschillende mogelijkheden en aspecten in hun samenhang tonen.

Om boodschappen vanuit een apparaat ten gehore te brengen hebben we nodig: een geheugen in welke vorm dan ook en een omzetter om de gecodeerde spraakgegevens die in het geheugen zijn opgeslagen weer in hoorbare signalen terug te brengen. Over het geheugen zullen we niet veel zeggen: het zou een tape kunnen zijn, maar meestal is het een digitaal geheugen (ROM, RAM, floppy disk, enz.). De toe te passen omzetters kunnen we globaal in een drietal groepen onderverdelen:

a. Golfvormcodering. Deze kunnen boodschappen reproduceren die van tevoren zijn opgenomen en gecodeerd en waarvan de golfvorm volgens een of ander recept is beschreven. Dat kan zijn PCM, waarvoor toch zo'n 64 kbit/sec. nodig is, tot aan de andere kant van de schaal LPC met multipuls-excitatie waarbij met 9600 bit/sec. al zeer goede spraakkwaliteit kan worden bereikt. Bij deze manier van opslaan van de spraak, is

het achteraf, bij het ten gehore brengen ervan, niet meer mogelijk wijzigingen in de boodschap aan te brengen (Voor de verschillende methodes van spraakcodering zie Deprettere, deze uitgave).

b. Resynthese van geanalyseerde spraak. Hierbij worden spraakboodschappen van tevoren opgenomen en geanalyseerd om er een parametrische beschrijving van te maken. Bij het ten gehore brengen van de zo opgeslagen spraakboodschappen moet men de klanken weer op grond van die parametrische beschrijving 'terug opbouwen' (= resynthetiseren). Het voordeel van deze methode is dat naast een aanzienlijke reductie van de benodigde bitrate bij de resynthese de spraakklanken nog gewijzigd kunnen worden (door namelijk vóór resynthese een of meerdere van die parameters te wijzigen). Dit is van groot belang om woorden of andere gebruikte fragmenten aan te passen aan de omgeving van de zin waarin ze zijn geplaatst. Dat geldt voor de duur van de klanken en vooral ook voor de toonhoogte. Het is gemakkelijk aan te tonen dat een dergelijke aanpassing de natuurlijkheid van de geproduceerde spraak aanzienlijk kan verbeteren.
We zullen de resynthese van spraak uitvoerig behandelen in paragraaf 5.

c. Spraaksynthese door regels. Bij deze vorm van spraaksynthese gaat men niet uit van van tevoren opgenomen spraak, maar de spraakboodschap wordt op basis van de tekst of fonetische tekst volledig kunstmatig gemaakt. Meestal gebeurt dat door een gering aantal kleine eenheden achter elkaar te schakelen. Men moet dan regels hanteren om de overgangen van de gebruikte eenheden op de juiste wijze te laten verlopen, om de gemaakte spraak zo natuurlijk mogelijk te laten klinken. Daarnaast moet ook de bovengenoemde aanpassing van de duuropbouw en het verloop van de toonhoogte plaatsvinden. Ook op deze methode van spraaksynthese door regels zullen we nader ingaan in paragraaf 7.

Een tweede doorsnijding van het spraaksynthesegebied heeft te maken met de toepassingen. Er is nogal wat verschil tussen een sprekende thermometer en sprekende telefoongids wat betreft het te kiezen systeem, de benodigde geheugenruimte enz.

a. Vaste boodschappen. Er zijn een aantal toepassingen waarin men gebruik maakt van een beperkt aantal vaste boodschappen. Enkele voorbeelden hiervan zijn: waarschuwingen in de auto, zoals 'Opgelet! Uw Oliepeil is te laag. Ga onmiddellijk naar een garage', of de bovengenoemde sprekende thermometer voor een blinde: 'Het is', 'negentien' 'graden'.

b. Variabele boodschappen. Er zijn toepassingen waarin de te geven boodschappen kunnen worden samengesteld uit korte fragmenten zoals woorden en woordgroepen, maar waarbij de fragmenten nog moeten worden aangepast aan de omgeving waarin ze voorkomen. Voorbeelden zijn: een sprekende klok die zegt: 'Het is nu' 'twintig' 'uur' 'dertien' of die kan zeggen 'Het is nu' 'dertien' 'uur' 'zeven'. In deze twee zinnen zal het woord 'dertien' verschillend klinken afhankelijk van de plaats in de zin. Een ander voorbeeld is het gesproken weerbericht of weerpraatje. Deze kunnen worden samengesteld met een betrekkelijk gering aantal woorden, echter ook hier is het nodig de woorden aan te passen aan de plaats in de zin, plaats van de klemtoon, enz.

c. Willekeurige tekst uitspreken. Dit zijn toepassingen waarin men geen van tevoren opgenomen spraak kan gebruiken, omdat óf de geheugenruimte niet toereikend is (sprekende encyclopedie, sprekend telefoonboek) óf omdat de spraakboodschappen nog niet vastliggen (geavanceerde informatiedialogen, spreekhulpmiddelen voor spraakgestoorden).

Tenslotte zijn er nog een drietal grootheden, die onderling afhankelijk zijn en die een belangrijke rol spelen bij de keuze van de een of andere oplossing voor een bepaald spraakoutputprobleem. Deze zijn: de complexiteit, de benodigde bitrate en de spraakkwaliteit.

a. De complexiteit van een codeer- of syntheseschakeling bepaalt vaak de prijs van het uiteindelijke apparaat, maar hangt nauw samen met de benodigde bitrate en dus ook met de grootte van het geheugen.

b. De benodigde bitrate hangt op zijn beurt weer heel sterk samen met de bereikte kwaliteit van de geproduceerde spraak. De uiterste grenzen waarbinnen de bitrate zal liggen zijn: aan de hoge kant ongeveer 100 kbit/sec. (of meer) en aan de lage kant ca 100 bit/sec. (Deze lage grens kan men afschatten door te bedenken dat er 40 verschillende spraakklanken zijn en dat per seconde zo'n 10 à 15 verschillende klanken door een spreker worden gezegd. Dan komt men ongeveer tot 100 bit/sec. informatie).

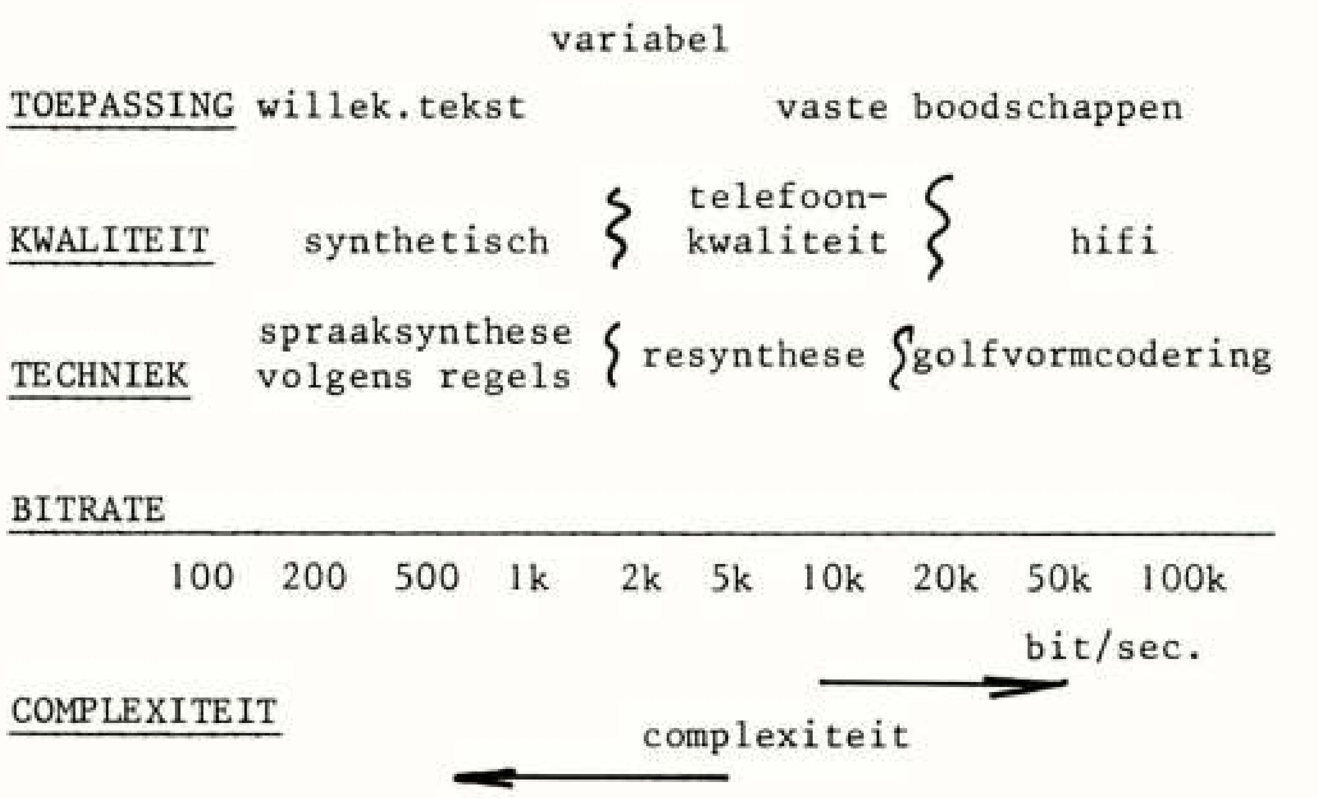


Fig. 1: Overzicht van verschillende grootheden, die in paragraaf 2 zijn besproken.

c. De spraakkwaliteit is natuurlijk een belangrijke eigenschap van een systeem. Er zijn geen objectieve methoden om de spraakkwaliteit te meten. Door middel van meestal tijdrovende luisterproeven kan men spraakkwaliteit kwantificeren (Steeneken, deze uitgave). Overigens is in de loop der jaren de spraakkwaliteit bij een bepaalde bitrate steeds toegenomen. De vooruitgang op dit gebied komt dus tot uitdrukking in óf een lagere bitrate óf een hogere spraakkwaliteit.

In Fig. 1 is getracht de hier genoemde aspecten in beeld te brengen.

3. NATUURLIJKE SPRAAK

Men kan zeggen dat het spraakgeluid wordt gevormd door een veranderlijke geluidsbron en een veranderlijk akoestisch filter dat het brongeluid wijzigt.
Voor de stemhebbende klanken (klinkers en een aantal medeklinkers als: m, n, l, b, d) ontstaat het brongeluid doordat de stembanden trillen. Deze trilling wordt veroorzaakt door een luchtdruk in de longen, die de stembanden uit elkaar duwt; dan gaat er lucht stromen; hierdoor ontstaat t.g.v. het Bernouilli-effect tussen de stembanden een onderdruk, waardoor de stembanden weer dichtgaan, daarbij ook nog geholpen door veerkracht in de stembanden. Hierdoor ontstaan luchtdrukimpulsen met een zekere herhalingsfrequentie. De bronfrequentie bepaalt de waargenomen toonhoogte. Een spreker regelt de bronfrequentie en dus de toonhoogte d.m.v. de mechanische spanning in de stembanden. De luidheid van de spraak wordt voornamelijk bepaald door het luchtdrukverschil tussen onder en boven de stembanden. Het filter voor de stemhebbende klanken is de mond- en keelholte. Als nasale klanken (m en n) worden gemaakt bestaat het filter ook nog uit de neusholte, omdat het zachte verhemelte het neuskanaal opent. Tijdens het spreken verandert voortdurend het mondkanaal van vorm, door bewegingen van de tong, kaak, enz. en dus verandert de filterwerking en daarom ook de klankkleur van het spraakgeluid.

Voor de stemloze wrijfklanken (f, s en g) is het brongeluid ruis die ontstaat door turbulentie van de luchtstroom uit de longen door een vernauwing in het mondkanaal. Voor de v en de z zijn er twee geluidsbronnen: trillende stembanden en luchtturbulentie. Het akoestisch filter bij deze klanken wordt gevormd door de holtes vóór en achter de vernauwing.
Bij plofklanken wordt het mondkanaal gedurende 50 ms tot 100 ms volledig afgesloten en dan weer geopend. Door de plotseling vrijkomende lucht wordt gedurende een korte tijd een ruisgeluid gevormd. In tegenstelling tot deze stemloze plofklanken (p, b, k) blijven bij de stemhebbende plofklanken (b, d) tijdens de afsluiting de stembanden juist doortrillen. Het akoestisch filter bij plofklanken wordt gevormd door de holtes vóór en achter de afsluiting.

4. SYNTHETISCHE SPRAAKKLANKEN

Bij het nabootsen van spraakklanken kan men ook een geluidsbron gevolgd door een filter nemen om zo spraakgeluid te vormen. In dit bron-filter-model wordt de bron U gevolgd door twee filters: het filter O gevormd door de keel- en mondholte en het filter R, dat de straling van het geluid bij de mondopening beschrijft (zie Fig. 2).

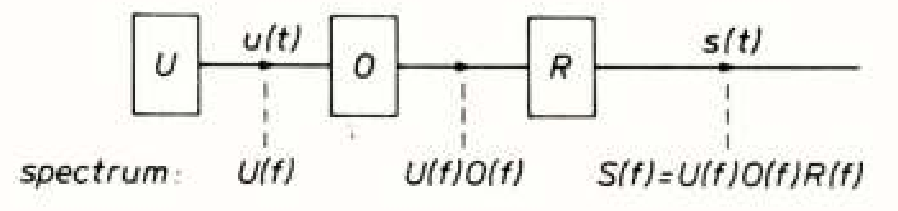


Fig. 2: Blokdiagram van het bron-filter-model.

Het brongeluid U is ofwel een reeks pulsen met een zekere herhalingsfrequentie ofwel ruis.
De overdrachtsfunctie O(f) is voornamelijk verantwoordelijk voor de klankkleur van het geluid. De mondkeelholte is te beschouwen als een wat grillig gevormde buis, die aan een kant -bij de stembanden- vrijwel gesloten is en aan de andere kant open. De overdrachtsfunctie van een dergelijke buis vertoont pieken bij de resonantiefrequenties. Deze pieken noemt men formanten. Elke formant wordt gekarakteriseerd door een middenfrequentie en een bandbreedte. Voor de waarneming van spraak zijn in het algemeen niet meer dan vijf formanten in het gebied tussen 0 Hz en 5000 Hz van belang. Deze worden over het algemeen aangeduid met F1 t/m F5.

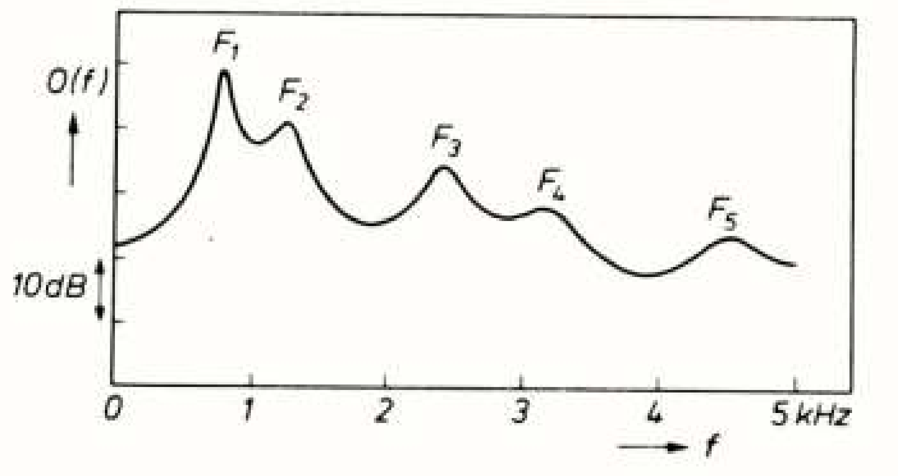


Fig. 3: Overdrachtsfunctie O(f) van een bepaalde mondstand met de formanten F1 t/m F5.

Apparaten of algorithmen voor spraaksynthese kan men baseren op dit bron-filter-model (zie Fig. 4). Als brongeluid neemt men ofwel periodieke impulsen met een zekere herhalingsfrequentie ofwel witte ruis. Dit brongeluid krijgt de gewenste sterkte door volume-instelling en wordt vervolgens gefilterd door een filter O'(f). In de overdrachtskarakteristiek van O' zijn verdisconteerd de veranderlijke eigenschappen van de mondkeelholte en verder de constante eigenschappen van de straling bij de mondopening (R in Fig. 2) en constante spectrale eigenschappen van de geluidsbron.

Voor stemhebbende signalen zijn in Fig. 4 enkele signalen met bijbehorende spectra geschetst.

Men zal bij het proces van spreken de mondstand steeds veranderen en dus zal ook het synthesemodel voortdurend veranderende parameters krijgen toegestuurd die het brongeluid en de overdrachtskarakteristiek bepalen. De snelheid waarmee de articulatoren bewegen is beperkt en dus kan men ook de sturende grootheden voor het synthesemodel ook met een overeenkomstig langzame snelheid veranderen. Dit is dan ook de reden waarom men een dergelijke parametrische beschrijving van het spraaksignaal met een geringere informatiestroom kan beschrijven dan het microfoonsignaal.

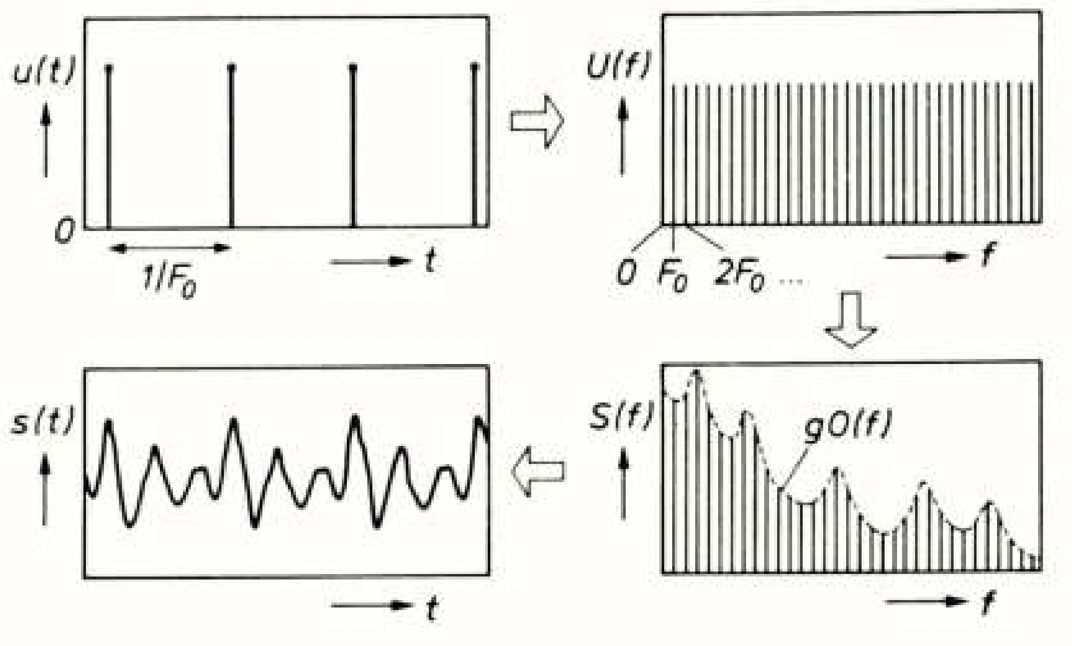


Fig. 4: Signalen en spectra in het synthesemodel voor stemhebbende klanken. Het brongeluid is u(t): periodieke deltapulsen met herhalingsfrequentie F_0. Het spectrum U(f) krijgt door het filter O'(f) de juiste spectrale samenstelling. Tenslotte is S(t) het gemaakte spraaksignaal.

5. SPRAAKRESYNTHESE

Het is mogelijk om de sturende grootheden voor zo'n synthesemodel uit natuurlijke spraak te bepalen. Op de analysemethoden zullen wij hier niet ingaan. In Fig. 5 is een compleet analyseresultaat getekend voor een Nederlandse zin gesproken door een mannenstem.
De analyse wordt 100 keer per seconde uitgevoerd, zodat een analyseresultaat beschikbaar is voor elke 10 ms. Deze frequentie voor het herhalen van de analyse is gebleken voldoende te zijn om het veranderende spraaksignaal te bemonsteren. De analyse wordt uitgevoerd over een spraaksegment van ongeveer 30 ms. In de bovenste twee hokken in Fig. 5 zijn de gegevens voor de geluidsbron weergegeven. De sterkte van het geluid G en de herhalingsfrequentie F0 van de stemhebbende geluidsbron. Tussen de hokken is nog aangegeven wanneer de ruisbron moet worden gebruikt.

In de onderste rechthoek zijn de gegevens geschetst die nodig zijn om het variabele filter in te stellen. Voor elk tijdstip (van 10 ms) worden de middenfrequenties van 5 formanten gegeven met bijbehorende kwaliteitsfactor. Met behulp van deze parametrische beschrijving is

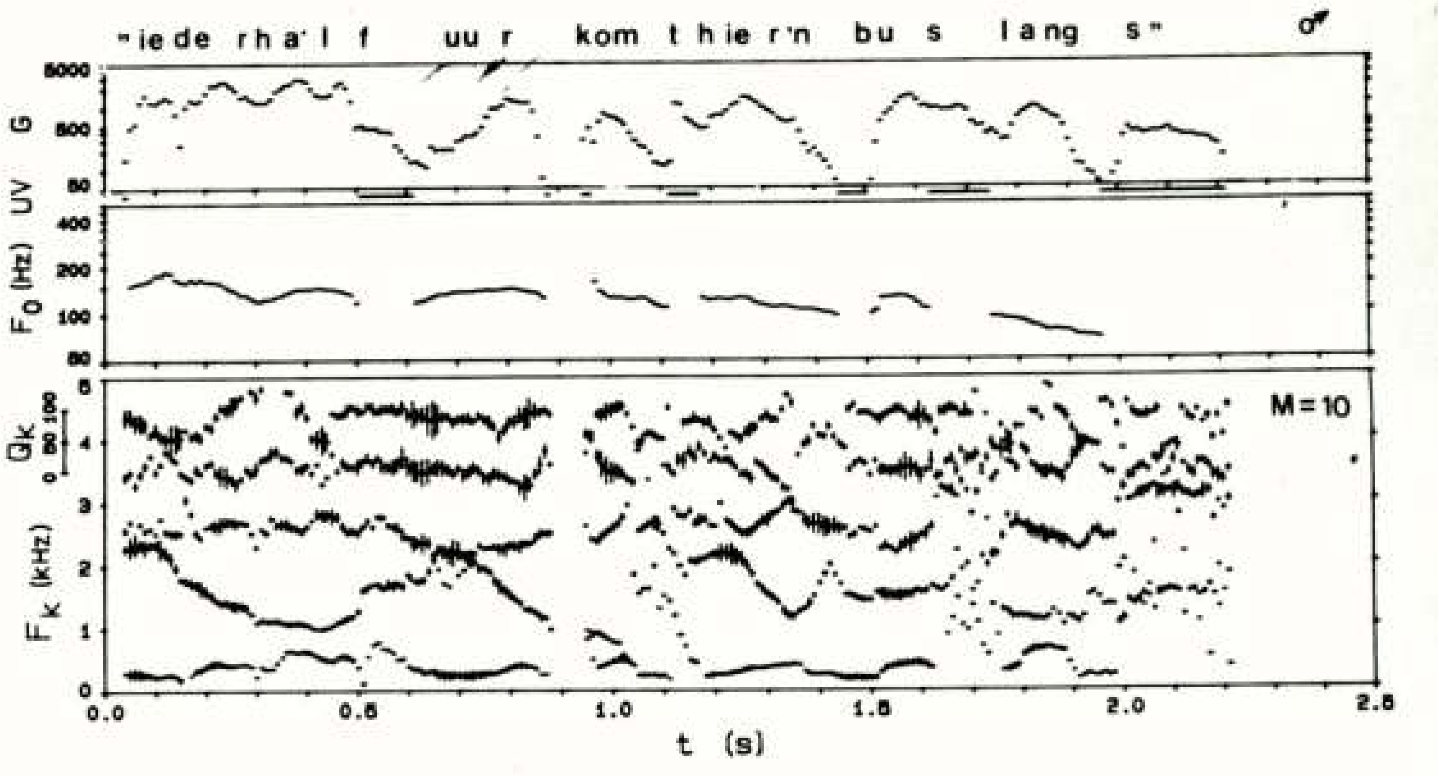


Fig. 5: Compleet analyseresultaat voor een mannenstem. Verklaring in de tekst.

het mogelijk heel behoorlijk spraak te resynthetiseren. Ook is het mogelijk om de parametrische beschrijving voor resynthese te wijzigen, bijvoorbeeld wat betreft de F0 (verantwoordelijk voor de waargenomen toonhoogte) en wat betreft de duur van spraaksegmenten. Dit was immers van groot belang om de geresynthetiseerde boodschappen natuurlijk te laten klinken.

Het variabel filter in het synthesemodel kan op verschillende wijzen geïmplementeerd worden: bijv. als ladderfilter of als spectrumshaper met bandfilters en amplituderegeling voor elk kanaal (zoals in kanaalvocoders). Het is echter bekend dat de hier gebruikte codering m.b.v. formanten de zuinigste beschrijving is. Een nadeel is echter dat de bepaling van de formanten uit natuurlijke spraak niet zonder problemen is.

Het verlies aan spraakkwaliteit dat men kan beluisteren bij deze spraakresynthese is te wijten aan het feit dat het bron-filter-model niet in staat is om de akoestische verschijnselen van het proces van spreken voldoende nauwkeurig te beschrijven. Zo zal het functioneren van de stembanden niet onafhankelijk zijn van de mondkeelholte. Ook is het gebruikte filter met een aantal resonantiepieken niet in staat de akoestische invloed van het neuskanaal te beschrijven of de invloed van de holtes achter de afsluiting bij wrijfklanken. Ook bij de bepaling van de verschillende grootheden gaat men ervan uit dat gedurende het analyse-interval (ca 30 ms) het signaal stationair is. Deze aanname zal zeker niet gelden bij plofklanken en andere snelle veranderingen.

6. SPRAAKCHIPS

Als men zo'n parametrische beschrijving heeft gemaakt, kan men met luisterexperimenten nagaan of op de codering van de gegevens kan worden bezuinigd. Eerst door de nauwkeurigheid waarmee elk gegeven wordt vastgelegd te beperken en ten tweede door de frequentie te beperken waarmee

de gegevens door nieuwe worden vervangen. Men kan nog verstaanbare spraak resynthetiseren met een bitrate van ongeveer 1000 bits/sec.

De laatste jaren hebben verschillende fabrikanten spraaksynthesechips gemaakt en op de markt gebracht, waarop een complete spraaksyntheseschakeling, meestal in digitale techniek, is ondergebracht. Ik zal hier een spraakchip: de MEA8000 van Philips, nader beschrijven die gebaseerd is op de al eerder beschreven codering in formanten. Het blokschema van de MEA8000 is weergegeven in Fig. 6. De codering voor deze chip is weergegeven in de onderstaande tabel I.

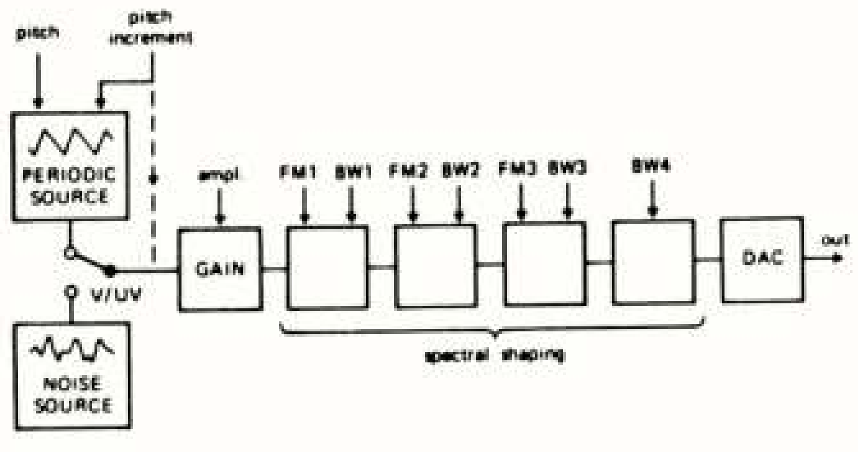


Fig. 6: Blokschema van de MEA8000 spraaksynthesechip.

Tabel I.

Afkorting	bits	parameter
FD	2	spraakframe duur (8, 16, 32, 64 ms)
AM	4	amplitude in log eenheden
PI	5	toename toonhoogte en ruis-keuze
F1	5	frequentie van formant 1
F2	5	frequentie van formant 2
F3	3	frequentie van formant 3
B1	2	bandbreedte van formant 1
B2	2	bandbreedte van formant 2
B3	2	bandbreedte van formant 3
B4	2	bandbreedte van formant 4
Totaal	32	

De frequentie van de vierde formant is vastgelegd op 3500 Hz. De frameduur wordt ook gecodeerd en met 2 bits kan men kiezen tussen 8 ms, 16 ms, 32 ms en 64 ms. Hieruit volgt dat de hoogste bitrate welke aan deze chip kan worden toegevoerd 4000 bits/sec. is (alle frameduren 8 ms) en de laagste bitrate is 500 bits/sec. (alle frameduren 64 ms). Het is de bedoeling om de frameduur aan te passen aan de mate waarmee het spraaksignaal zelf verandert: bij een snelle overgang gebruike men korte segmenten en in stabiele stukken gebruike men lange segmenten). In de chip worden de grootheden 8 keer per frame geïnterpoleerd om zodoende grote overgangen (die zeker bij lange frameduren zouden optreden) glad te strijken. In de praktijk liggen de benodigde bitrates voor goed verstaanbare spraak tussen de 1000 en 2000 bits/sec. In een toepassing van zo'n spraakchip heeft men naast deze chip ook nog nodig een geheugen (PROM of ROM) waarin de gecodeerde spraak ligt opgeslagen en een microprocessor die het datatransport regelt. Voor een toepassing zal men een aantal boodschappen of fragmenten van meldingen (denk aan een sprekende klok) van tevoren door een spreker laten zeggen, laten analyseren door een computer of spraakontwikkelsysteem (kan door de fabrikant van de chip worden gedaan) en tenslotte in een geheugen laten vastleggen. Er zijn intussen een groot aantal van dergelijke spraaksynthesechips te koop. De toepassing ervan komt echter traag op gang.

7. WILLEKEURIGE TEKST UITSPREKEN

Wil men willekeurige teksten laten uitspreken door een automaat, dan moet de tekst eerst omgezet worden in een fonetische transcriptie om vervolgens door een spraaksynthese-door-regels-systeem te worden omgezet in verstaanbare spraak. Het eerste probleem: de omzetting van tekst in een fonetische transcriptie beschouw ik hier als gegeven (zie Boot, deze uitgave). Ik ga ook ervan uit dat de fonetische transcriptie is voorzien van indicaties waar lettergrepen klemtoon krijgen.
Bij het tweede probleem, dat van het spraaksynthese-door-regels-systeem staat centraal de vraag uit welke eenheden zal men de spraakuiting samenstellen. Neemt men weinig eenheden, zoals de elementaire spraakklanken (soms fonemen genoemd) dan heeft men er slechts weinig nodig (ca 40), maar de regels die nodig zijn om vervolgens de klanken aan te passen aan hun omgeving zullen nogal ingewikkeld zijn. Vooral de overgang van de ene klank naar de andere is moeilijk met behulp van regels te beschrijven. Neemt men daarentegen grote eenheden bijv. woorden dan is het duidelijk dat men zeer veel geheugenruimte nodig heeft voor de opslag, maar dat de regels voor aanpassing aan de omgeving veel simpeler zullen zijn.

Een aardig compromis, dat de laatste jaren nogal wat aandacht krijgt, lijkt te zijn difoon-synthese. De eenheden zijn difonen: stukje klank + overgang + stukje volgende klank. Daardoor heeft men de overgangen niet door regels hoeven te beschrijven en het bovengenoemde probleem is zodoende omzeild. Het aan elkaar koppelen van spraaksegmenten in de meer stabiele stukken geeft vrijwel geen problemen. Voor een dergelijk systeem heeft men ca 1600 difonen nodig.
In een systeem dat door ons gebouwd wordt, waarvoor de input is: fonetische tekst met klemtoontekens en de genoemde spraakchip het uitvoerorgaan is, wordt ongeveer 50 kbyte gebruikt voor de opslag van de difonen. De codering van de spraakgegevens voor de difonen is dezelfde als in paragraaf 5 is beschreven. In Fig. 7 is de codering geschetst van het woord 'banaan', samengesteld uit difonen. Op de difoongrenzen, waar de fragmenten aan el-

kaar gekoppeld zijn, kan men kleine discontinuïteiten zien, maar men kan ze vrijwel niet horen.

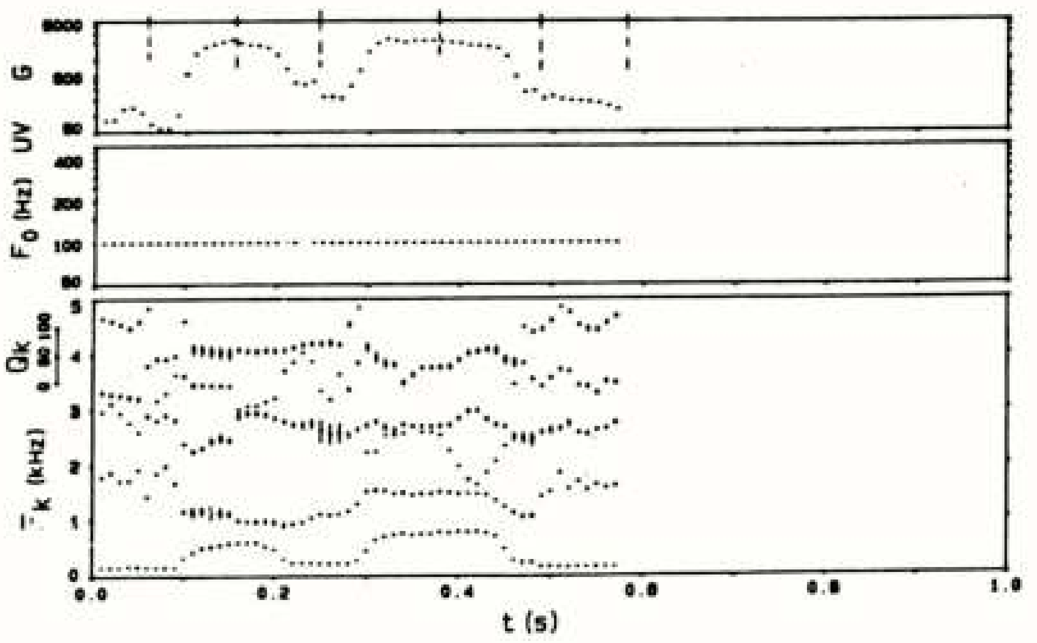


Fig. 7: Parametrische beschrijving voor het woord 'banaan', verkregen door difoonconcatenatie. De difoongrenzen zijn aangegeven met stippellijnen.

De gegevens voor de difonen zijn gehaald uit beklemtoond uitgesproken lettergrepen uit onzinwoorden als 'nenaane'. Hieruit kan men het difoon 'naa' en het difoon 'aan' halen. Heeft men nu een zin samengesteld uit dergelijke difoonfragmenten dan klinkt zoiets nog helemaal monotoon. Een grote sprong in natuurlijkheid krijgt men door de toonhoogte aan te passen aan de intonatie van een dergelijke Nederlandse zin. Ook zal aanpassing van de duren van de segmenten aan de plaats in de zin verbetering geven. Immers de difonen zijn allemaal gehaald uit beklemtoonde lettergrepen en ze komen in een zin ook voor op niet beklemtoonde posities.

8. SLOTOPMERKINGEN

Het spraakonderzoek krijgt tegenwoordig nogal wat aandacht. Dit zal onder andere ertoe leiden dat de kwaliteit van synthetische spraak steeds zal verbeteren. Ik wil hier enkele mogelijkheden noemen, die er zijn om het proces van spreken nauwkeuriger in kaart te brengen en zodoende de kunst van het opwekken van synthetische spraak vooruit te helpen.

- Verlaten van bron-filter-model. De generator van het brongeluid (de stemband-oscillator) wordt onafhankelijk beschouwd van het akoestische filter (het mondkanaal). De aannames die hierin worden gemaakt vormen een te grote beperking. Ingewikkelder modellen betekenen echter ook complexere synthesetechnieken en moeizamere analysemethoden.
- De aanname van de (quasi-)stationariteit vormt ook een grote beperking. Er zijn te veel spraaksegmenten, die hierdoor niet of slecht worden weergegeven in de analyseresultaten.
- Er is nog betrekkelijk weinig kennis omtrent de juiste duuropbouw van spraakuitingen. Dit komt o.a. tot uiting in de difoonconcatenatie.
- Een groot probleem, dat wel aandacht begint te krijgen, maar toch nog niet opgelost is, is de fonetische transcriptie of anders gezegd de grafeem-foneem-omzetter.

De vele aandacht voor spraak zal ook tot uiting komen in meer toepassingen dan tot nu toe zijn gemaakt. Bekend zijn: 'Speak and Spell' van Texas Instruments dat een zekere pioniersrol heeft vervuld en voorts het sprekende dashboard van een type Renault.
Dat er nog ruimte voor eenvoudige toepassingen is blijkt wel uit het feit dat bij de landing van de eerste space shuttle de ene hooggetrainde piloot aan de andere piloot de stand van de hoogtemeter moest voorlezen.

Ik ben van mening dat de toepassingen van spraaksynthese pas goed op gang zullen komen, als de apparaten ook onze spraak kunnen verstaan, zodat er een natuurlijke dialoog mogelijk is tussen de mens en de machine.

Tijdens de voordracht werd een en ander met geluidsvoorbeelden geïllustreerd.

Voor verdere lezing aanbevolen:

Flanagan, J.L. and Rabiner, L.R. (eds). Speech synthesis. Benchmark Paper in Acoustics.

Hart, J. 't et al. Manipulaties met spraakgeluid. Philips Technisch Tijdschrift 40, no. 4, 108-119.

MEA8000 voice synthesizer: principles and interfacing. Techn. Publication 101, Elcoma.

Witten, Ian H. Principles o- computer speech. 1982. Academic Press.

Enkele artikelen in Databus n2 7/8, juli/augustus 1982.

Voordracht gehouden tijdens de 319e werkvergadering.

HARDWARE VOOR SPRAAKCODERINGSSYSTEMEN: 'SPECIAL-PURPOSE'- EN 'GENERAL-PURPOSE'-CHIPS.

Ir. G.J. Bosscha
Philips' Natuurkundig Laboratorium
5600 JA Eindhoven

On the hardware for speech coding systems: special pupose and general purpose chips. Hardware realisations of a harmonic-sieve pitch extractor and a residual-excited coder (9.6 kbit/s) on the basis of special purpose and general purpose chips respectively are presented. For both realisations the performance is discussed.

1. INLEIDING

Een hardware-realisatie van een spraakcoderingssysteem wordt tegenwoordig in vrijwel alle gevallen voorafgegaan door uitgebreide computersimulaties. In eerste instantie gaat het hierbij om de functionele aspecten van het coderingssysteem, waarbij de volledige nauwkeurigheid van de computer wordt benut; pas in tweede instantie gaan effecten ten gevolge van eindige woordlengte in de simulaties een rol spelen, om een afbeelding naar hardware te kunnen realiseren. Deze afbeelding kan eenduidig gemaakt worden, zodat resultaten (zoals uiteindelijke spraakkwaliteit) verkregen uit de simulaties identiek zijn aan die uit de hardware. In deze simulaties moet dan wel rekening gehouden worden met de eventuele beperkingen die de hardware oplegt. Hiervoor is een goede kennis van de digitale signaalbehandeling en van de huidige digitale technologie vereist.

Bij de keuze van de hardware voor een systeem (of een deelfunctie ervan) doen zich ruwweg twee mogelijkheden voor: 'special-purpose'-chips ('maatwerk') of 'general-purpose'-chips ('confectie'). In het eerste geval moeten er één of meer geïntegreerde circuits worden ontworpen en in het tweede geval wordt gebruik gemaakt van bestaande programmeerbare digitale signaalprocessoren. Criteria voor deze keuze zijn o.a. complexiteit van het systeem, het aantal te realiseren systemen, kosten per systeem, ontwerptijd en de mogelijkheid tot het aanbrengen van modificaties in een later stadium.

Op enkele van deze criteria en op voor- en nadelen van beide hardware-mogelijkheden wordt in paragrafen 3 en 4 nader ingegaan. Voor elke mogelijkheid gebeurt dit aan de hand van een voorbeeld uit de (laboratorium-)praktijk, te weten een toonhoogtemeter ('pitch extractor') gebaseerd op het principe van de harmonische zeef en een residu-coder werkend op een bitsnelheid van 9,6 kbit/s. Alvorens naar deze voorbeelden over te gaan, wordt in de volgende paragraaf nader ingegaan op enkele principes die eraan ten grondslag liggen.

2. PRINCIPES

Spraakcoderingssystemen kunnen in drie klassen worden onderverdeeld: golfvormcoders, vocoders of parametrische coders en hybridische of gemengde coders (zie fig.1),(Sluyter, 1984). In het algemeen geldt dat de spraakkwaliteit toeneemt met de bitsnelheid, indien voor elke bitsnelheid de optimale coderingstechniek wordt gekozen.

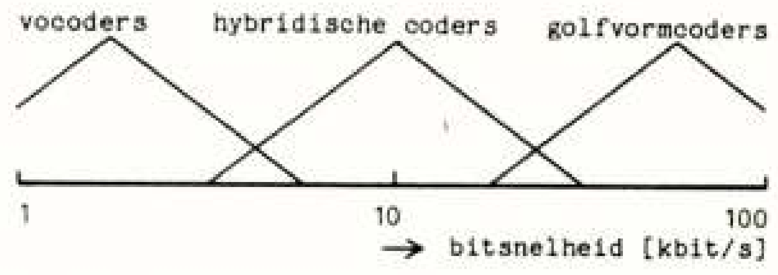


Fig. 1. Klassificatie van spraakcoderingssystemen

Een voorbeeld van golfvormcodering is Puls-Code-Modulatie, die een goede spraakkwaliteit geeft bij een bitsnelheid van 64 kbit/s.

Voor bitsnelheden lager dan ca. 5 kbit/s moeten vocoders worden toegepast. Vocoders zijn gebaseerd op een spraakproduktiemodel (zie fig.2). Het spraaksignaal wordt ontleed in excitatie-parameters (toonhoogte en beslissing stemhebbend/stemloos) en in parameters die het stemkanaal beschrijven. De spraaksynthese is gebaseerd op het spraakproduktiemodel waarin het synthesefilter het stemkanaal modelleert. Dit filter

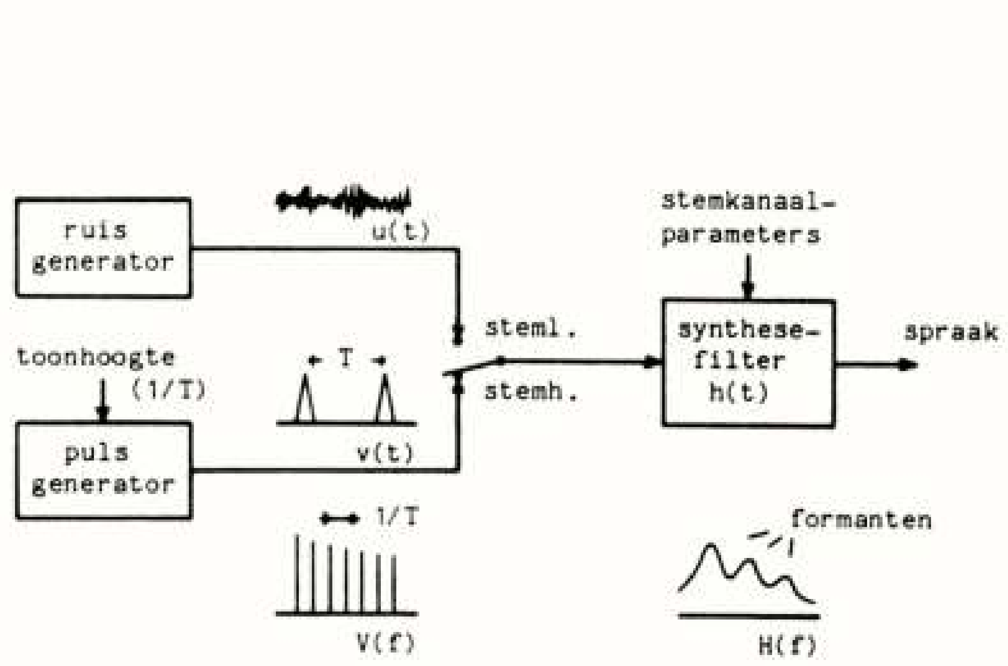


Fig. 2. Spraakproduktiemodel

wordt voor stemhebbende spraak geëxciteerd door een periodieke pulsreeks met de juiste toonhoogte en voor stemloze spraak door witte ruis. Het spectrum van stemhebbende spraak zal dus een harmonische structuur hebben.

Hybridische coders maken gebruik van zowel golfvormcoder- als vocoder-technieken.

Het eerstvolgende hardware-voorbeeld dat behandeld wordt betreft een deelfunctie van een vocoder nl. een toonhoogtemeter, die voorzien is van een stemhebbend/stemloos-detector en die voor stemhebbende spraak de grondtoon bepaalt, en het tweede een compleet coderingssysteem nl. een hybridische coder met een bitsnelheid van 9.6 kbit/s.

3. 'SPECIAL-PURPOSE'-CHIPS

De toonhoogtemeter gebaseerd op het principe van de harmonische zeef (Zuidweg, 1982) kan met de volgende drie 'special-purpose'-chips worden gerealiseerd: een segmenteringsbuffer (SEB), een processor die de discrete Fourier-transformatie uitvoert (DFT) en een 16bits-microcomputer (MIC) (zie fig.3). Het segmenteringsbuffer splitst na analoog/digitaal-omzetting (8kHz bemonsteringsfrequentie) het spraaksignaal op in (overlappende) segmenten van 256 monsters (32 msec). Van elk segment berekent de DFT-processor het complexe spectrum in de band van 0-2kHz. De microcomputer voert aan de hand van dit spectrum de algoritmes voor stemhebbend/stemloos-detectie en voor grondtoonextractie uit, resulterend in de excitatie-parameters.

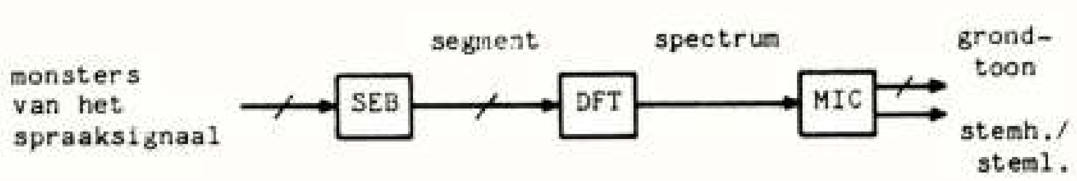


Fig. 3. LSI-systeem voor de harmonische-zeef-grondtoon-extractor

De chips bevatten resp. 46000 (SEB), 20000 (DFT) en 35000 (MIC) transistoren en zijn verpakt in resp. een 40-, 28- en 40-pins behuizing. Elk verbruikt bij een klokfrequentie van 3,648 MHz een vermogen van ongeveer 0.4 Watt.

In de volgende paragrafen wordt elk van deze chips in wat meer detail beschreven.

3.1 Segmenteringsbuffer

Het segmenteringsbuffer bestaat uit twee 12bits-brede schuifregisters IB en SB met een lengte van 96 respectievelijk 256 woorden (zie fig.4). Gedurende elke spectrumberekening van de DFT-processor worden de waarden uit het circulerend schuifregister SB 128 maal aan die processor aangeboden, terwijl nieuw binnenkomende monsters in het circulerend schuifregister IB worden opgeslagen (schakelaar S2:'a'; S1:'a' & 'b', schakelaar S1 wordt zodanig bediend dat in IB een opeenvolgende reeks van monsters ontstaat). Zodra de spectrumberekening gereed gekomen is, worden de waarden uit IB doorgeschoven naar SB (S2:'b'). Nieuwe monsters van het spraaksignaal worden nu rechtstreeks in SB opgeslagen (S2:'c'), zodat voor een nieuwe spectrumberekening steeds het meest recente spraaksegment beschikbaar is.

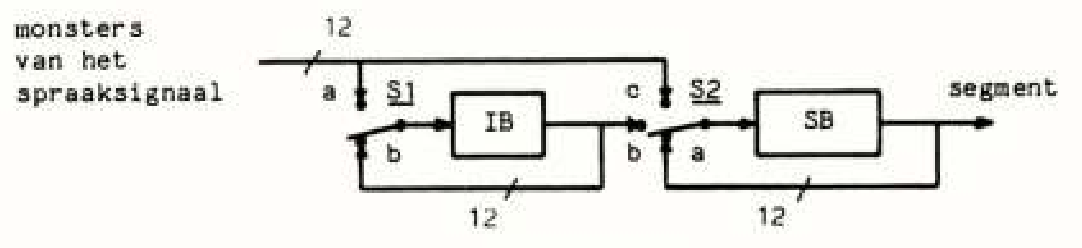


Fig. 4. Segmenterings-buffer

3.2 DFT-processor

De DFT-processor (van Meerbergen, 1983) voert de volgende algoritme uit:

$$Re\{F(k)\}=\sum_{i=0}^{255} s(i)w(i)\cos(2\pi ik/256)$$

$$Im\{F(k)\}=\sum_{i=0}^{255} s(i)w(i)\sin(2\pi ik/256)$$

voor frequentiepunten:

$k=0,1,2,...63$ (band 0-2kHz) of

$k=64,65,..127$ (band 2-4kHz).

Hierin stellen $Re\{F(k)\}$ en $Im\{F(k)\}$ het reële en imaginaire deel van het spectrum $F(k)$ voor, $s(i)$ de monsters van het spraaksegment (afkomstig van het segmenteringsbuffer) en $w(i)$ een weegfunctie:

$w(i)=0.54-0.46\cos(2\pi i/256)$, ('Hamming window').

Een blokschema van deze processor is weergegeven in figuur 5. De cosinus/sinus- en de weegfunctie-waarden zijn opgeslagen in twee (256*8)bits-ROMs. Deze waarden worden in een parallelle (8*8)bits-vermenigvuldiger verwerkt, waarna het resultaat op 12 bits wordt afgebroken en vermenigvuldigd met de monsters uit een spraaksegment in een parallelle (12*12)bits-

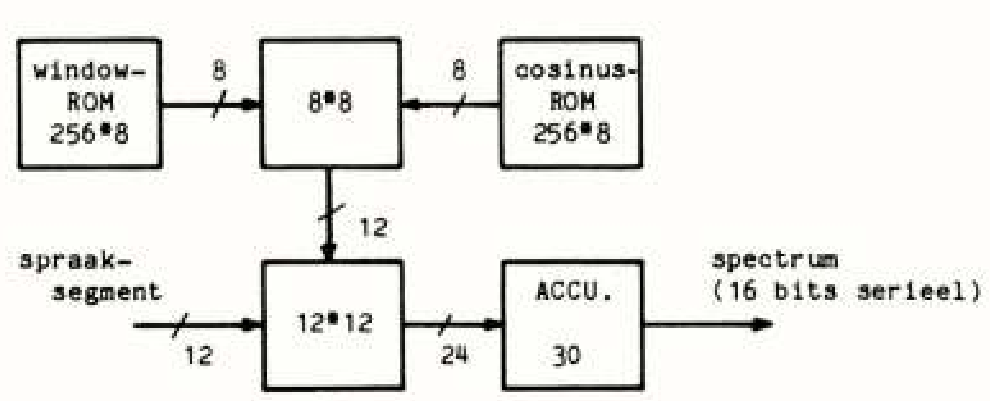


Fig. 5. DFT-processor

vermenigvuldiger. Deze resultaten worden opgeteld bij de inhoud van een 30bits-accumulator, waarna aan het einde van elke cyclus de meest significante 16 bits (van Re{F(k)} of Im{F(k)}) naar de microcomputer worden gezonden. Naast het berekenen van het spectrum in de band van 0-2kHz of in de band van 2-4kHz kan ook gestart worden op een willekeurig frequentiepunt, dat serieel moet worden ingelezen. Het is mogelijk om verschillende DFT-processoren parallel te schakelen om de totale rekentijd te verminderen en/of het gehele spectrum (0-4kHz) in eenmaal te berekenen.

3.3 Microcomputer

De microcomputer berekent uit de reële en imaginaire waarden van het spectrum F(k), afkomstig van de DFT-processor, het amplitudespectrum. Daarna vindt een stemhebbend/stemloos-detectie plaats op basis van spectrale intensiteitsvariaties van opeenvolgende segmenten. In het stemhebbende geval worden de pieken in het spectrum ('spraakcomponenten') gelokaliseerd. Om de harmonische structuur ervan te bepalen, worden deze componenten vergeleken met een set van 63 harmonische patronen, elk met een andere grondtoon in het gebied van 50-500Hz (zie fig.6). Deze patronen kunnen worden voorgesteld als één-dimensionale zeven. Elke zeef vertoont een gat op de frequentie van de grondtoon en op een aantal harmonischen daarvan. Afhankelijk van de verhouding van het aantal spraakcomponenten dat elke zeef wel en niet kan passeren, wordt een zeef gevonden die het beste past. De grondtoon ervan geeft dan de toonhoogte aan (Sluyter, 1982).

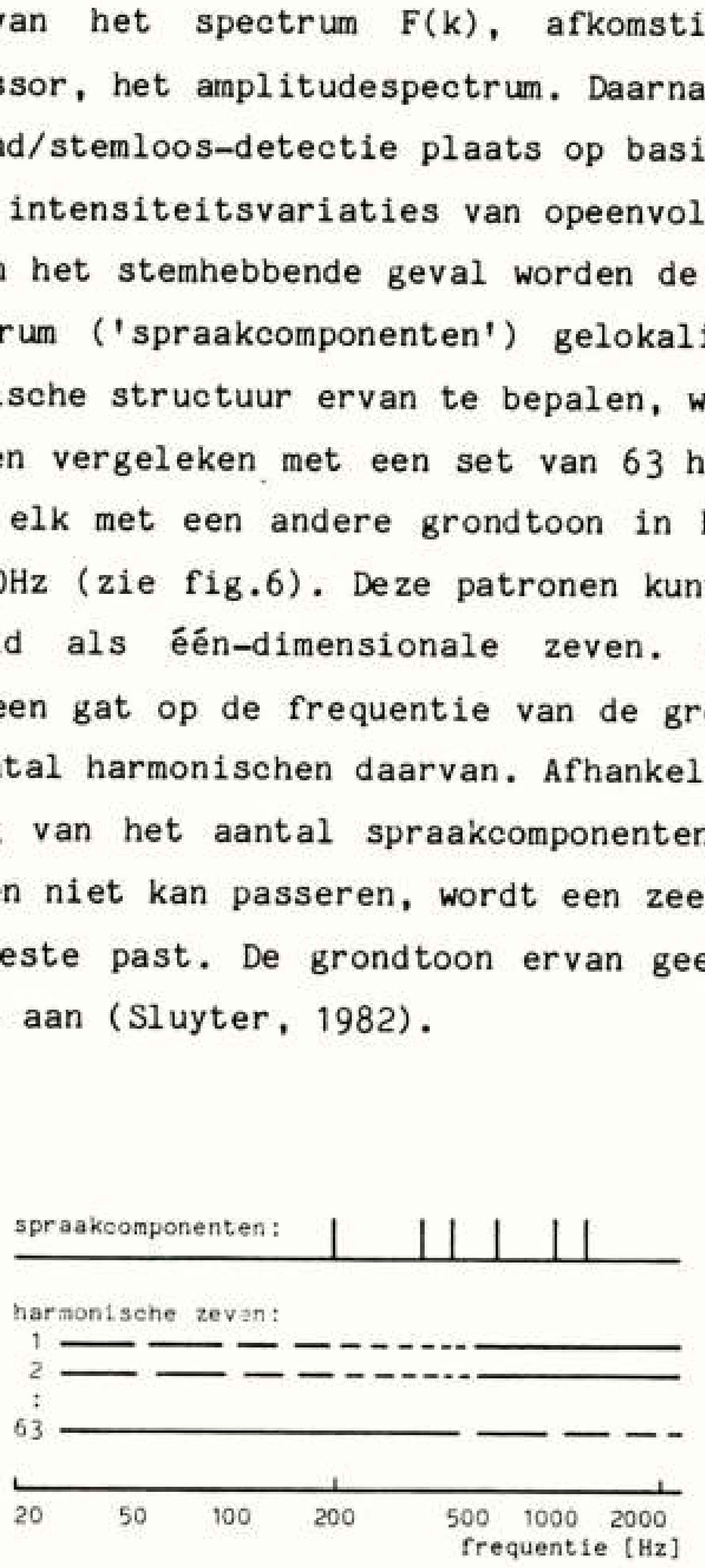


Fig. 6. Het principe van de harmonische zeef

Het vergelijken van de harmonische patronen met de spraakcomponenten vergt de meeste bewerkingen. Daarom is deze processor uitgerust met vier 'arithmetic/logic units', om de algoritme efficiënt uit te voeren. Verder kenmerkt deze chip zich door een 16bits-woordlengte voor de data, een datageheugen dat bestaat uit een (152*16)bits-RAM en een (104*16)bits-ROM, en een programmageheugen van (512*40)bits-woorden. De I/O-interface bestaat uit twee seriële ingangen, een seriële uitgang en een parallelle uitgang.

3.4 Eigenschappen

Bovengenoemde drie chips zijn specifiek voor de 'harmonische-zeef-grondtoon-extractor' ontworpen. De chip-set in totaliteit voert dan ook alleen deze functie uit. Hiervoor bevat de microcomputer een vast programma (masker geprogrammeerd), waarin het aanbrengen van veranderingen of van een nieuw programma tijdrovend en kostbaar is. Echter, de combinatie van segmenteringsbuffer met één of meer DFT-processoren is breder toepasbaar, daar de DFT-processor op verschillende manieren kan worden ingesteld (spectrum van 0-2Khz, 2-4kHz en de spectrale waarden voor een enkel frequentiepunt). Momenteel wordt deze combinatie toegepast in o.a. een DFT-vocoder (Bosscha, 1982) en een woordherkenningssysteem voor losse woorden (Geppert, 1984).

Enkele voordelen van deze chip-set zijn: een relatief klein volume, een laag vermogensverbruik en (bijna) geen additionele hardware voor besturingscircuits en dergelijke. Enkele nadelen van deze aanpak zijn: lange ontwerptijd en weinig flexibiliteit ten aanzien van modificaties gedurende het ontwerptraject of erna. Zodra met het ontwerp van 'special-purpose'-chips een begin wordt gemaakt dienen de specificaties van het systeem bevroren te worden. Verder moet de IC-technologie en moeten de bijbehorende CAD-hulpmiddelen adequaat en betrouwbaar zijn.

De kosten van 'special-purpose'-chips zijn hoog vanwege o.a. de inzet van mensen uit verschillende disciplines en de dure technologie. Uit kostenoverweging zullen deze chips dan ook alleen aantrekkelijk zijn als er grote aantallen van worden omgezet.

4. 'GENERAL-PURPOSE'-CHIPS

In deze paragraaf wordt eerst ingegaan op de werking van de residu-coder, daarna komt de hardware voor deze coder aan de orde. Ook zal blijken dat deze hardware toepasbaar is voor spraakcoderingssystemen in het algemeen.

4.1 Residu-coder

Het principe van de residu-coder is weergegeven in figuur 7. In de analyse volgens lineaire predictie (LPC) wordt periodiek een verzameling van parameters bepaald, die een benadering vertegenwoordigen van de spectrale omhullende van het spraaksignaal. Het predictieve filter wordt aan de hand van deze parameters zodanig ingesteld, dat zijn overdrachtsfunctie A(z) gelijk wordt aan de inverse van die spectrale omhullende. Na filtering van het betreffende spraaksignaal ontstaat een residu-signaal dat een vrijwel vlakke

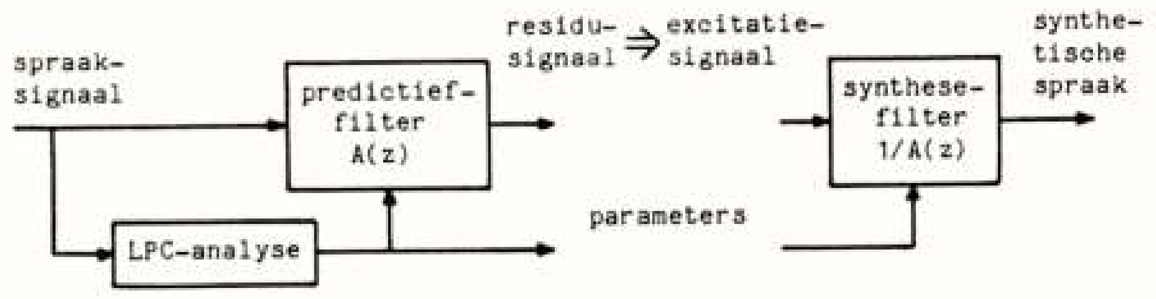


Fig. 7. Residu-coder

spectrale omhullende heeft. Dit signaal kan dienen als excitatie-signaal voor het synthesefilter, dat met dezelfde parameters wordt ingesteld maar dan met overdrachtsfunctie 1/A(z). Daar beide filters elkaars inverse zijn, wordt weer het originele spraaksignaal verkregen. Echter, van enige reductie van de bitsnelheid is nog geen sprake. Door nu het spectraal vlak zijn van het residu-signaal te benutten, kan de bitsnelheid tot ca. 9,6 kbit/s worden gereduceerd met behoud van een goede spraakkwaliteit (Sluyter, 1983).

4.2 Parallelle verwerkingsconfiguratie

Bovengenoemde spraakcoder is gerealiseerd met programmeerbare digitale signaalprocessoren van NEC, type µPD7720 (Product description, 1981). Deze processor onderscheidt zich door een snelle en efficiënte verwerking van data, omdat ze is uitgerust met flexibele in- en uitgangspoorten (zowel parallel als serieel) en met een parallelle (16*16)bits-vermenigvuldiger. De tijd benodigd voor een instructie bedraagt 250 nsec en het vermogensverbruik ca. 1 Watt. Verder bevat het een datageheugen dat bestaat uit een (128*16)bits-RAM en een (510*13)bits-ROM, en een programmageheugen van (512*23)bits-woorden. Het geheel is verpakt in een 28pins behuizing.

De algoritme van de spraakcoder is te complex om door één processor te worden uitgevoerd. Voor deze toepassing is het geheugen ervan te klein en de verwerkingssnelheid te laag. Om deze tekortkomingen te omzeilen is gekozen voor een parallelle verwerkingsconfiguratie (zie fig.8). Deze configuratie bestaat uit een aantal modules (met elk een signaalprocessor SP, een busarbiter BA en een datacontroller DC), een analoge en digitale I/O, en een gemeenschappelijke timing. De signaalprocessoren in deze modules kunnen

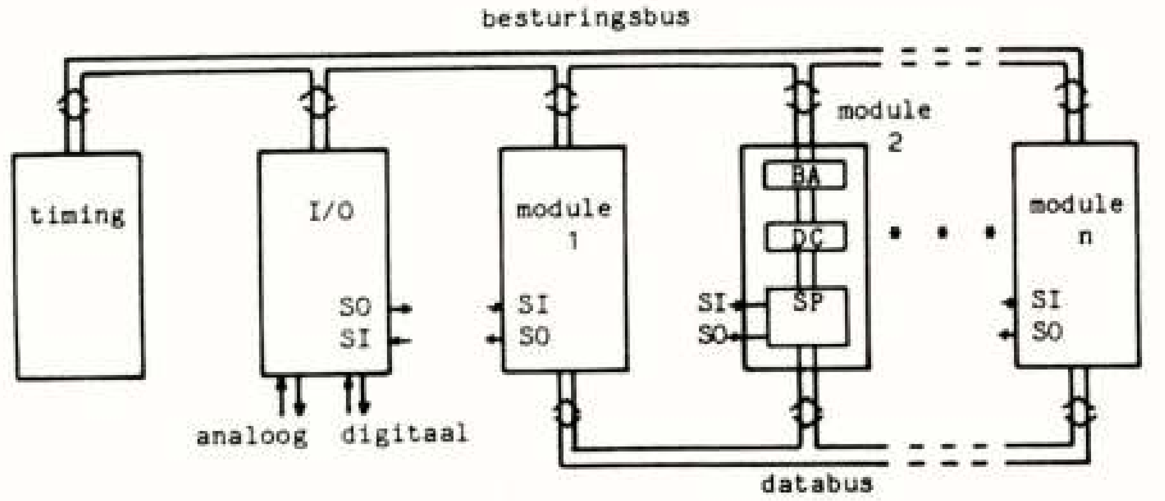


Fig. 8. Parallelle verwerkingsconfiguratie

onderling communiceren via een 8bits-brede databus, waarbij de busarbiter het busverkeer regelt en de datacontroller gedurende het datatransport voor de besturingssignalen zorgt.

De communicatie tussen de processoren wordt volledig door de software bepaald. Hiervoor bevat elke processor een zendroutine en een ontvangroutine. Indien een processor een dataverzameling wil gaan verzenden, vraagt deze via zijn busarbiter om toegang tot de bus. Als deze vrij is, of zodra deze vrijkomt, krijgt de processor de bus, waarna deze processor er een datacode op plaatst. Op basis van deze code wordt een processor geselecteerd die de bijbehorende dataverzameling wil ontvangen. Zodra die processor gereed is voor ontvangst, meldt hij dat via zijn datacontroller aan die van de zender, waarna beide processoren gesynchroniseerd worden. Hierdoor kunnen woorden uit de dataverzameling, snel en efficiënt, achter elkaar worden overgeheveld (ca. 1,25 usec per 16bits-woord).

De analoge en digitale I/O zorgt voor de communicatie van de seriële in- en uitgangen op de processoren (SI en SO) naar de buitenwereld. De analoge I/O kan bestaan uit analoog/digitaal- en digitaal/analoog-omzetters, de digitale I/O uit een interface voor aansluiting op modems.

De timing voor de modules is gemeenschappelijk en is onafhankelijk van het te realiseren systeem. De timing daarentegen die betrekking heeft op de analoge en digitale I/O, zoals bijvoorbeeld rasterperiode, bitsnelheid en bemonsteringsfrequentie, is systeemafhankelijk.

De hiervoor beschreven spraakcoder is gerealiseerd met in totaal zes modules (processoren) waarvan er voor de zender vier gebruikt zijn en voor de ontvanger twee.

4.3 Eigenschappen

Aangezien de communicatie tussen de modules onderling volledig door de software bestuurd wordt, is de hierboven beschreven parallelle verwerkingsconfiguratie voor verschillende digitale systemen toepasbaar. Alleen de specifieke timing voor de I/O-module moet voor elk systeem worden aangepast.

Enkele voordelen van deze configuratie zijn: een korte ontwerptijd voor de hardware, grote flexibiliteit ten aanzien van modificaties tijdens en na het ontwerptraject en een redelijk klein volume. Een nadeel kan zijn het relatief hoge vermogensverbruik.

De beperking van deze configuratie wordt gevormd door de mate waarin algoritmes kunnen worden opgesplitst, zodat ze over de verschillende processoren kunnen worden verdeeld. Dit opsplitsen kan niet willekeurig ver worden doorgevoerd, aangezien dan de databus overbezet raakt.

Op dit moment vergt het vertalen van software waarmee het systeem is gesimuleerd (in een hogere programmertaal, zoals bijvoorbeeld FORTRAN) naar software voor de processoren (microprogramma's) relatief veel tijd. Een compiler die deze vertaalslag efficiënt uitvoert zou dit proces zeer versnellen (een aanzet hiertoe wordt beschreven in (Hermann, 1983)).

5. CONCLUSIES

Aan de hand van twee in hardware gerealiseerde systemen, namelijk een deelfunctie van een vocoder: een toonhoogtemeter volgens het principe van de harmonische zeef, en een compleet spraakcoderingssysteem: een residu-coder met een bitsnelheid van 9,6 kbit/s, zijn verschillende eigenschappen van zowel 'special-purpose'- als 'general-purpose'-chips beschreven.

Als compactheid, laag vermogensverbruik of grote aantallen chips een rol spelen, zal in de regel gekozen worden voor een systeem op basis van 'special-purpose'-chips. De ontwerptijd ervan kan lang zijn, afhankelijk van de complexiteit.

In die gevallen waarin een snelle realisatie van belang is, kan een digitaal (spraakcoderings-)systeem in een relatief korte tijd, met behulp van signaalprocessoren in een parallelle verwerkingsconfiguratie, verwezenlijkt worden. De meeste ontwerptijd hierbij gaat zitten in het schrijven van microprogramma's voor de signaalprocessoren. Hierin weerspiegelt zich nu een tendens dat de tijd benodigd voor een hardware-ontwerp afneemt en dat die voor software-ontwikkeling toeneemt.

Bovendien kan met deze configuratie snel worden ingespeeld op de volgende generatie van signaalprocessoren die naar alle waarschijnlijkheid eind '84 of begin '85 op de markt zal komen. Deze zal zich onderscheiden door een meer efficiënte architectuur, een grotere geheugencapaciteit, een snellere gegevensverwerking en een lager vermogensverbruik. Ook zullen er compilers komen die de software-ontwikkeling voor deze processoren eenvoudiger maken. Dan zullen ook complexere systemen snel en efficiënt gerealiseerd kunnen worden.

6. LITERATUUR

G.J. Bosscha and R.J. Sluyter, 'DFT-vocoder using harmonic-sieve pitch extraction', IEEE Int. Conf. on ASSP, pp.1952-1955, may 1982, Paris.

R. Geppert and P. Schartau, 'A DFT-based front-end for word recognition systems', wordt gepubliceerd op IEEE Int. Conf. on ASSP, march 1984, San Diego.

O.E. Hermann and J. Smit, 'A user-friendly environment to implement algorithms on single-chip digital signal processors', Proceedings EUSIPCO, september 1983, Erlangen.

J.L. van Meerbergen and F.J. van Wijk, 'A 2µm NMOS 256-point discrete Fourier transform processor', IEEE Int. Solid-State Circ. Conf., pp. 124-125, 1983, New York.

N.E.C. Microcomputers, 'µPD7720 Signal Processing Interface (SPI) User's Manual'.

R.J. Sluyter, H.J. Kotmans and T.A.C.M. Claasen, 'Improvements of the harmonic-sieve pitch extraction scheme and an appropriate method for voiced-unvoiced detection', IEEE Int. Conf. on ASSP, pp. 188-191, may 1982, Paris.

R.J. Sluyter, 'Spraakcodering voor het mobiele radiokanaal', K.I.V.I. Leergang TH-Eindhoven, Mobiele Communicatie, april 1983, Eindhoven.

R.J. Sluyter, 'Digitalisering van spraak', verschijnt binnenkort in Philips Technisch Tijdschrift, jaargang 41, no. 7/8.

P.Zuidweg, J.L. van Meerbergen and M.L. van der Meulen, 'Custom LSI chip-set for speech analysis', IEEE Int. Conf. on ASSP, pp. 521-524,may 1982, Paris.

Voordracht gehouden tijdens de 319e werkvergadering.

EVALUATIE VAN SPRAAKPRODUKTIE- EN SPRAAKHERKENNINGSSYSTEMEN

H.J.M. Steeneken
Instituut voor Zintuigfysiologie TNO
Kampweg 5
3769 DE Soesterberg

1. INLEIDING

Het toepassen van spraakproduktie- en spraakherkenningssystemen zal in dit decennium een grote vlucht nemen. Voor wat betreft automatische spraakherkenning (ASH) zijn hiervoor een aantal oorzaken te noemen:

- onderzoek inzake codering van spraak (vocoders) heeft tot technieken geleid die goed toepasbaar zijn bij ASH,
- dynamische programmering heeft de benodigde normalisatie in duur tussen overeenkomstige spraaksignalen mogelijk gemaakt,
- moderne signaalprocessoren en microcomputers maken het mogelijk de herkenning binnen zeer korte tijd uit te voeren.

Op dit moment zijn dan ook enige tientallen systemen op de markt beschikbaar. Ook de ontwikkeling van spraakproduktiesystemen heeft door de moderne digitale signaalbehandelingstechnieken een grote ontwikkeling doorgemaakt en er zijn vele systemen commercieel beschikbaar. Dit zal tot gevolg hebben dat deze systemen veelvuldig zullen worden toegepast met name om de relatie mens-machine te vereenvoudigen. We kunnen hierbij denken aan in- en uitvoer van gegevens bij computers, post-, giraal- en telefoonverkeer maar ook aan voorzieningen voor motorisch gehandicapten, bij het besturen van een rolstoel, of andere aktiviteiten in het dagelijks leven. Voor elke toepassing zal echter moeten worden nagegaan welk systeem hier het best bij aansluit. Dit heeft zowel met de prijs als met specifieke eigenschappen van het systeem te maken.

De evaluatie van vooral automatische spraakherkenningssystemen staat echter nog in de kinderschoenen. Er zijn geen gestandaardiseerde methoden om deze systemen onderling te vergelijken. Wel wordt er gewerkt aan een internationaal bestand van testmateriaal (vaste testwoordenlijsten) om tot een uniforme testmethode te komen. Dit is echter specialistenwerk; er zal ook moeten worden nagegaan of er geen eenvoudige, meer algemeen toepasbare, meetmethoden kunnen worden ontwikkeld.

In deze voordracht zullen we een aantal meetmethoden beschrijven en nagaan door welke faktoren de resultaten kunnen worden beïnvloed. Omdat deze vraag voor produktiesystemen een totaal andere aanpak vereist dan bij automatische spraakherkenningssystemen zullen we deze twee onderwerpen afzonderlijk behandelen.

2. SPRAAKPRODUKTIE

2.1 Keuze van een spraakproduktiesysteem

Bij alle methoden van spraakproduktie wordt gebruik gemaakt van spraaksignalen die zijn voortgebracht door de menselijke stem. Moderne technieken hebben het mogelijk gemaakt de spraaksignalen efficiënter op te slaan. In Fig. 1 wordt hiervan een voorbeeld gegeven.

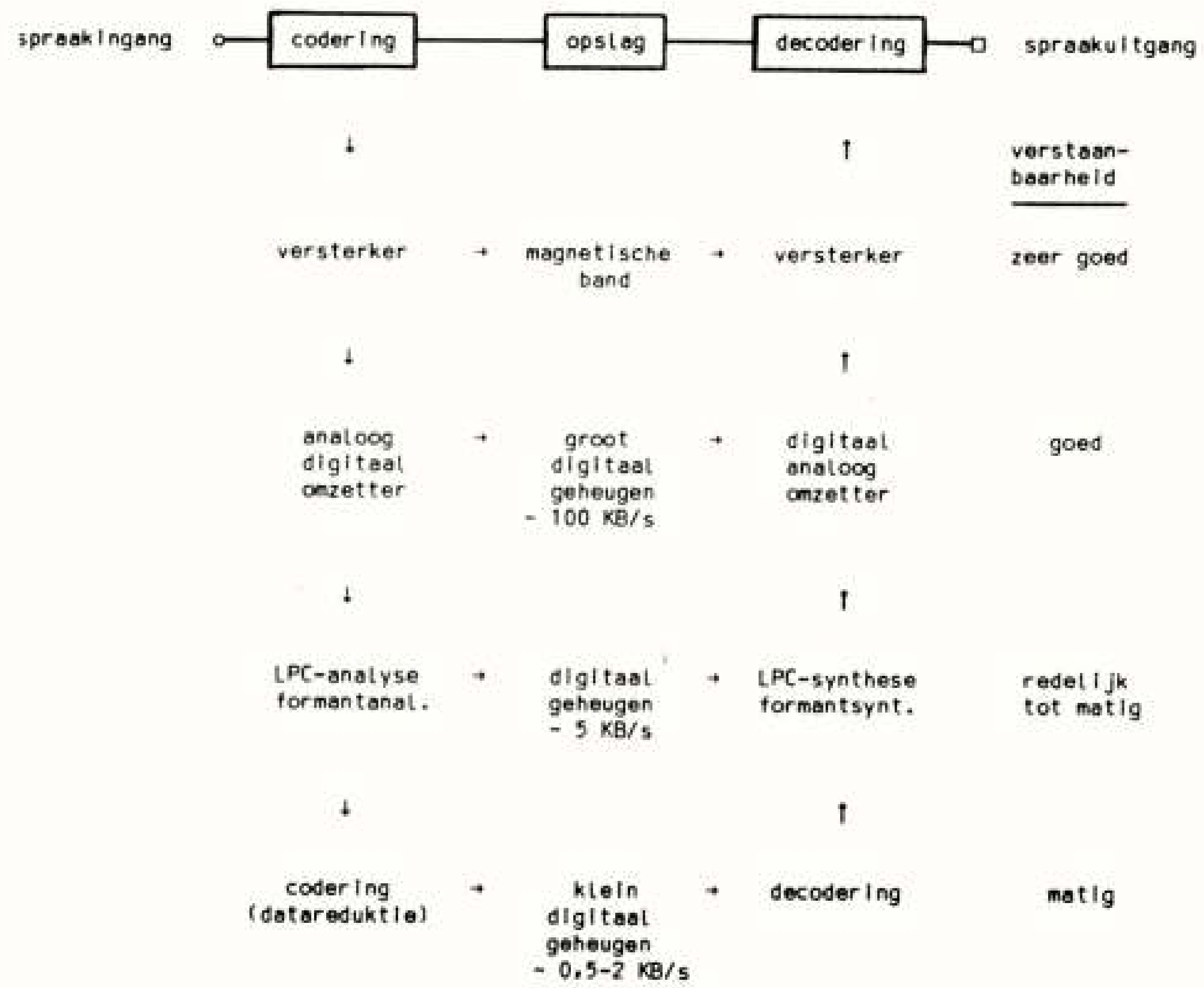


Fig. 1 Principe van enige spraakanalyse- en spraakproduktiemethoden met de benodigde digitale opslag in bits per sekonde en de te verwachten verstaanbaarheid.

Bij het hier weergegeven principe wordt, na een eventuele codering, het spraaksignaal opgeslagen waarna de signalen op elk gewenst moment en in elke volgorde kunnen worden weergegeven (bij registratie op magnetische band is dit niet mogelijk). Bij opslag in een digitaal geheugen, zoals computer, disk of elektronisch geheugen, dient de golfvorm van het spraaksignaal eerst te worden gedigitaliseerd via een analoog-digitaal omzetter. Voor het nauwkeurig vastleggen van de golfvorm van een spraaksignaal zijn ca. 10.000 bemonsteringen per sekonde nodig en dient de amplitude met een nauwkeurigheid van ca. 1000 niveaus (10 bit) te worden gedigitaliseerd. Men heeft dan voor een sekonde signaal een geheugen nodig van 100 kB. Er is daarom naar een techniek gezocht om deze grote hoeveelheid getallen te reduceren met behoud van zoveel mogelijk informatie. Hierbij wordt niet meer de golfvorm van het signaal vastgelegd, maar andere kenmerkende eigenschappen zoals

frekwentiespektrum, grondfrekwentie en signaalsterkte. Bij spraak kan dit optimaal gebeuren door formantanalyse of LPC technieken (linear predictive coding). De benodigde opslagruimte wordt hiermee beperkt tot ca. 5 kB/s. Verdere reduktie vindt plaats door van speciale eigenschappen van een spraaksignaal gebruik te maken, zoals bij een lang aangehouden klank (b.v. aa in aan) waarbij eenmalig alle codering voor de klank wordt opgeslagen tezamen met de benodigde duur. Ook is het mogelijk de fonemen die in de spraak voorkomen vast te leggen en deze vervolgens in de gewenste volgorde terug te genereren. Uit deze veelheid van systemen dienen we er een of meer te kiezen die bij de toepassing past om dan tot een eventuele evaluatie over te gaan.

2.2 Het bepalen van de kwaliteit van spraakproduktiesystemen

Over het algemeen worden spraakproduktiesystemen slechts kwalitatief geëvalueerd en worden er geen kwantitatieve metingen van de spraakverstaanbaarheid uitgevoerd. Toch hebben onderzoekingen naar de kwaliteit van spraakkommunikatiesystemen, vooral in de veertiger jaren, tot gevolg gehad dat vele subjektieve meetmethoden zijn ontwikkeld. Op dit moment zijn de meest toegepaste methoden: de logatoomtest, de diagnostische rijmtest, de "Mean Opinion Score" en het bepalen van de verstaanbaarheid met het spelalfabet of cijfers.

Elke test heeft zijn eigen voordelen en beperkingen. Het hangt meestal af van de faciliteiten die de onderzoeker tot zijn beschikking heeft, welke meetmethode wordt toegepast.

Voor het meten van de Mean Opinion Score (MOS) wordt gebruik gemaakt van eenvoudige zinnen; hierbij wordt aan een groep luisteraars gevraagd elke zin afzonderlijk te beoordelen ten aanzien van de verstaanbaarheid. Bij deze beoordeling heeft de luisteraar een schaal met meestal vijf intervallen ter beschikking (uitstekend, goed, matig, slecht, onmogelijk). De methode werkt snel en er kunnen ongetrainde luisteraars worden gebruikt. Vanwege de grove manier van beoordelen en beïnvloeding van de beoordeling door de beste en slechtste konditie die aan de luisteraars wordt aangeboden is de methode niet nauwkeurig.

De diagnostische rijmtest (DRT) is een twee-keuzetest waarbij de luisteraar op papier of beeldscherm visueel twee woorden krijgt aangeboden die slechts qua beginmedeklinker verschillen zoals mar - nar. Slechts een van deze woorden wordt uitgesproken en de luisteraar moet beslissen welk woord dat was. Bij het samenstellen van de DRT is ervoor gezorgd dat alleen kombinaties tussen medeklinkers worden vergeleken die slechts voor één fonetisch kenmerk verschillen. De in de DRT gebruikte fonetische kenmerken zijn: stemhebbend, nasaal, aangehouden, sisachtig, labiaal en tongafsluitend [1, 2]. Een DRT bestaat uit ca. 100 woordparen. Men kan met ongetrainde luisteraars werken. De meetnauwkeurigheid is groter dan bij de MOS.

Bij de logatoomtest wordt gebruik gemaakt van éénlettergrepige woorden, kombinaties van medeklinker-klinker-medeklinker, die over het algemeen geen betekenis hebben. Het voordeel hiervan is dat logatomen een redundantie gelijk nul bezitten en dus, in tegenstelling tot zinnen, alle individuele fonemen moeten worden waargenomen.

Een logatoomtest bestaat uit het aanbieden van een lijst van 50 logatoomwoorden, waarbij er voor gezorgd wordt dat de frekwentie van het voorkomen van de klinkers en begin- en eindmedeklinkers representatief is voor de (Nederlandse) taal. Bij een dergelijke test is het tevens mogelijk de afzonderlijke klinker- en medeklinkerverstaanbaarheid te bepalen.

De verschillen tussen de aangeboden fonemen en de door de luisteraar gehoorde fonemen kan worden vastgelegd in een verwarringsmatrix. Op basis van deze matrix kan een optimale woordenlijst worden samengesteld voor het toepassen van het systeem. De logatoomtest levert resultaten die zeer betrouwbaar zijn, maar vereist getrainde luisteraars.

Verstaanbaarheidstests worden over het algemeen uitgevoerd met meerdere sprekers en luisteraars omdat er grote individuele verschillen optreden. In Fig. 2 is voor vier verstaanbaarheidsmaten de verstaanbaarheid gegeven als funktie van de signaal-ruisverhouding van een spraakverbinding.

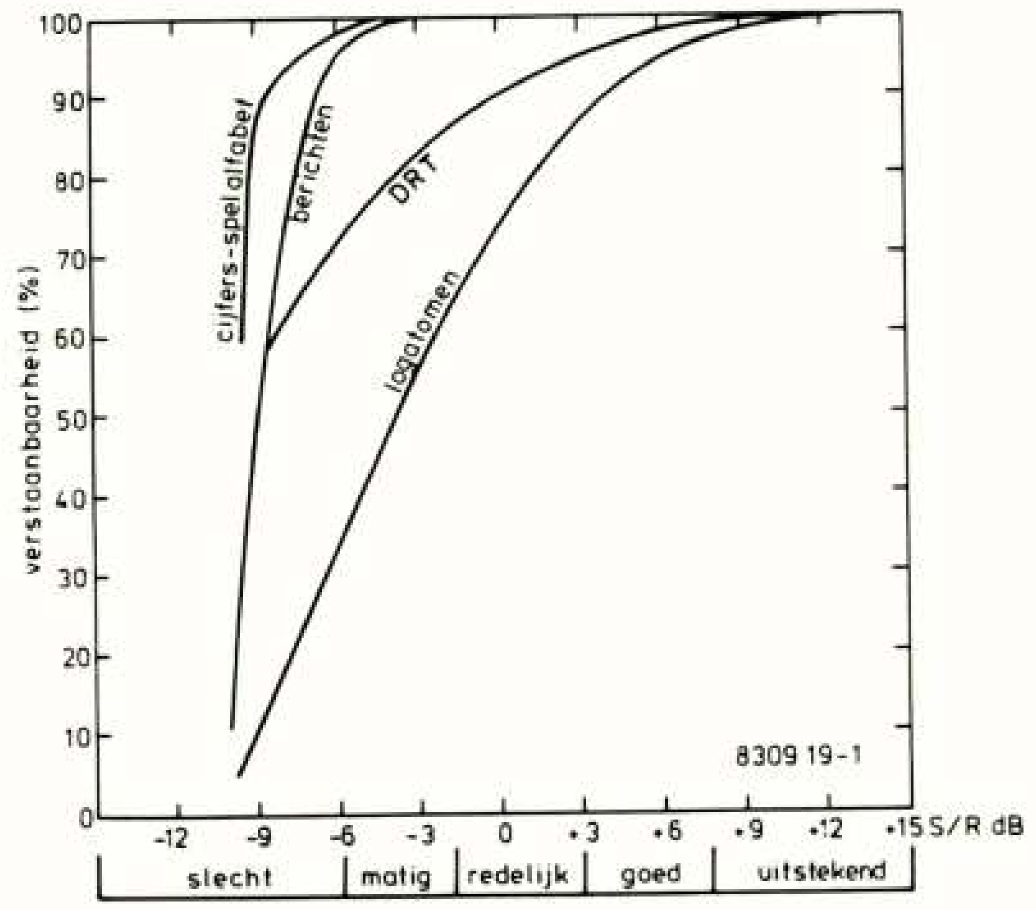


Fig. 2 Relatie tussen signaal-ruisverhouding en enige verstaanbaarheidsmaten.

De gegeven relaties tussen verstaanbaarheid en signaal-ruisverhouding gelden alleen voor stoorruis waarbij het frekwentiespektrum gelijk is aan het gemiddelde frekwentiespektrum van spraak (bijv. geroezemoes), en waarbij geen andere vervormingen aanwezig zijn.

Ter illustratie wordt in Tabel I voor twee spraakproduktiesystemen de logatoomverstaanbaarheid tezamen met de klinker- en de begin- en eindmedeklinkerverstaanbaarheid gegeven. Deze verstaanbaarheden zijn bepaald met één spreker en acht luisteraars, voor het Texas Instruments TMS 5220 en het Philips MEA 8000 systeem.

Tabel I Gemiddelde verstaanbaarheid in procent van logatomen en klinkers, begin- en eindmedeklinkers afzonderlijk voor twee spraakproduktiesystemen.

	TMS 5220	MEA 8000
logatomen	53,0	60,8
beginmedeklinkers	59,2	66,5
klinkers	97,5	97,0
eindmedeklinkers	87,0	91,0

In Tabel II is voor de Philips MEA 8000 de verwarringsmatrix voor de beginmedeklinkers gegeven zoals deze behoren bij de in Tabel I gegeven verstaanbaarheid van 66,5%.

Als we de logatoomverstaanbaarheid van beide systemen via Fig. 2 omzetten in een beoordeling blijkt dat bij de TMS 5220 van een lage verstaanbaarheid gesproken kan worden, terwijl met de MEA 8000 een matige verstaanbaarheid wordt bereikt. Deze resultaten gelden overigens voor een spreker en kunnen enigszins wijzigen voor meerdere sprekers. Voor beide systemen is de zinsverstaanbaarheid 100%. Tabel II toont aan dat de beginmedeklinker "B" veelal verwisseld wordt met de "W" en de "P"; de beginmedeklinker "F" met de "V". Hiermee kan bij het toepassen van dit systeem en deze spreker rekening worden gehouden.

Tabel II Verwarringsmatrix voor de beginmedeklinkers bepaald voor het spraakproduktiesysteem MEA 8000.

			verstaan																		
RESPONSIE			B	D	F	G	H	J	K	L	M	N	P	R	CS	T	V	W	Z	??	% goed
STIMULUS																					
	1	B	1	-	-	-	-	-	-	-	-	-	7	1	-	1	-	14	-	-	4,2
	2	D	-	56	-	-	-	1	-	-	-	-	-	-	-	6	-	1	-	-	87,5
	3	F	-	-	2	-	-	-	-	-	-	-	-	-	-	-	6	-	-	-	25,0
a	4	G	-	-	5	7	4	-	-	-	-	-	1	-	-	3	4	-	-	-	29,2
a	5	H	5	-	-	-	14	-	-	-	-	-	-	-	-	-	-	5	-	-	58,3
n	6	J	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	-	-	100,0
g	7	K	-	-	-	-	1	-	12	-	-	-	3	-	-	-	-	-	-	-	75,0
e	8	L	-	-	-	-	-	-	-	16	-	-	-	-	-	-	-	-	-	-	100,0
b	9	M	-	-	-	-	-	-	-	-	13	1	-	-	-	-	-	2	-	-	81,3
o	10	N	-	-	-	-	-	-	-	-	-	24	-	-	-	-	-	-	-	-	100,0
d	11	P	-	-	-	-	-	-	-	-	-	-	4	-	-	4	-	-	-	-	50,0
e	12	R	-	-	-	-	1	-	-	-	-	-	-	31	-	-	-	16	-	-	64,6
n	13	CS	-	-	-	-	-	-	-	-	-	-	-	-	14	-	-	-	2	-	87,5
	14	T	-	-	-	-	-	-	-	-	-	-	-	-	-	24	-	-	-	-	100,0
	15	V	-	-	-	-	8	-	-	-	-	-	1	-	-	-	13	10	-	-	40,6
	16	W	3	-	-	-	-	-	-	1	-	1	-	-	-	-	1	18	-	-	75,0
	17	Z	-	7	-	-	-	5	-	-	-	-	-	-	-	3	-	-	9	-	37,5

3. AUTOMATISCHE SPRAAKHERKENNING

3.1 Keuze van een spraakherkenningssysteem

Voorafgaand aan de keuze van een spraak- of woordherkennend systeem dient een inventarisatie te worden gemaakt om na te gaan of het installeren van een dergelijk systeem zinvol is. Hiervoor dient gelet te worden op kostenbesparing, stressvermindering bij de gebruikers en ergonomische aspekten. Voor de juiste keuze van een spraakherkennend systeem dienen de specifieke eigenschappen van dit systeem te korresponderen met de taak die door dit systeem moet worden uitgevoerd [3]. Enige belangrijke eigenschappen zijn:

soortherkenning	=	losse woorden, aaneengesloten woorden of lopend spraak
vocabulaire	=	aantal woorden dat moet worden herkend
training	=	soort referentiepatronen, aantal referenties, automatische bijstelling van de referenties
werkwijze	=	o.a. onafhankelijkheid van het signaalniveau,tijdnormering,gevoeligheid voor stoorlawaai, duur herkenning na invoer, sprekerafhankelijkheid
uitvoering	=	afmetingen, voeding, gewicht, prijs, aansluitmogelijkheden.

Om tot een optimale keuze te kunnen komen, maakt men meestal een inventarisatie van de werkplek door een mens-machine dialoog te konstrueren [4] waarbij ergonomische aspekten worden geïnventariseerd zoals:

- plaatsing mikrofoon
- omgevingslawaai
- spreeksnelheid
- grootte vocabulaire
- woordkeuze
- syntaxregels
- herkenningsnauwkeurigheid
- terugmelding bij akseptatie
- foutkorrektie
- stabiliteit referentiepatronen
- leereffekt gebruikers.

Op basis van deze gegevens kan een keuze worden gemaakt van potentiële voor de taak geschikte systemen. Een verdere evaluatie dient dan plaats te vinden op basis van de beschikbare specifikaties of door zelf de prestaties van de gekozen herkenners te meten.

Voor het eerste geval wordt door Lea [5] een methode aangegeven waarbij de benodigde specifikatie-items afzonderlijk worden gewogen om in relatie met de taak te worden gebracht. Op deze wijze kan een tabel worden gemaakt van specifikatie-items per systeem, en kan worden bepaald welk systeem of systemen het hoogst scoren. Een voorbeeld van deze methode is gegeven in referentie [5].

3.2 Het bepalen van de kwaliteit van een spraakherkenningssysteem

Bij het evalueren van een woordherkennend systeem is de meest voor de hand liggende vraag: "Welk percentage van de woorden wordt korrekt herkend?". Om deze vraag te beantwoorden moeten we ons tevens afvragen welke faktoren dit percentage beïnvloeden. We kunnen deze faktoren in een aantal, naar soort ingedeelde hoofdgroepen onderscheiden, met name: soort spraak, sprekereigenschappen, taak van de toepassing, akoestische omgeving van de invoerpositie, specifikatie van het ingangssignaal en herkennereigenschappen. Elk van deze hoofdgroepen is weer onder te verdelen in specifieke, afzonderlijke te onderzoeken, faktoren. Deze faktoren zijn:

- SPRAAK	geïsoleerde woorden aaneengesloten woorden lopende spraak
- SPREKER	sprekerafhankelijkheid spreker bekend/onbekend in referentiepatronen leeftijd, sexe, aksent, taal moment van opname spraakniveau spreeksnelheid taakafhankelijkheid
- TAAK	omvang redundantie vocabulaire syntaxkomplexiteit
- OMGEVING	achtergrondlawaai nagalm andere spraak
- INGANGSSIGNAAL	mikrofoon ruis vervorming/bandbreedte
- HERKENNER	systeemparameters, drempels etc.

Voor vele toepassingen is het niet voldoende het percentage herkende woorden te bepalen, maar dient ook te worden nagegaan of er verwisselingen bij de herkenning optreden en of woorden die buiten de toegepaste vocabulaire vallen, toch worden herkend. Voor herkenners van geïsoleerde woorden kunnen we de volgende prestatiematen bepalen:

- Binnen de vocabulaire	percentage herkend percentage niet herkend percentage verkeerd herkend
- Buiten de vocabulaire	percentage niet herkend percentage verkeerd herkend

- Optredende verwarringen tussen woorden binnen en buiten de vocabulaire
- Subjektief equivalent ruisniveau.

Hierbij spreken de herkenningspercentages, voor woorden binnen en buiten de vocabulaire van referentiewoorden, voor zichzelf. We kunnen uit de meetresultaten echter ook een verwarringsmatrix opstellen, vergelijkbaar met Tabel II, waaruit valt af te lezen welke antwoorden door de herkenner werden gegeven per aangeboden woord. Hiermee kan de mate waarin elk woord wordt verwisseld met een ander woord (bv. een - en) worden bepaald.

Voor herkenners van aaneengesloten woorden kan aan de lijst van prestatiematen de volgende worden toegevoegd:

- Percentage tussenvoegingen,
- Percentage weglatingen.

Bovenstaande prestatiematen zijn zeer afhankelijk van de gebruikte vocabulaire, aantal sprekers en dergelijke. We kunnen een meer algemene prestatiemaat, onafhankelijk van de vocabulaire, bepalen door na te gaan in hoeverre luisteraars met dezelfde vocabulaire tot eenzelfde score zouden zijn gekomen als aan de testwoorden een maskerend ruissignaal is toegevoegd. De sterkte van de ruis en het type woordmateriaal is dus bepalend voor het percentage. Het ruisniveau benodigd om tot eenzelfde percentage met luisteraars als met de woordherkenner te komen, wordt het equivalente ruisniveau genoemd. De relatie volgens Fig. 2 tussen de verschillende verstaanbaarheidsmaten als funktie van de signaal-ruisverhouding maakt een vergelijking met soortgelijke, elders uitgevoerde, metingen mogelijk. Om het vergelijken van verschillende herkenners te vereenvoudigen, zonder alle metingen te herhalen, wordt er internationaal naar gestreefd een genormaliseerde vocabulaire te gebruiken (data base). Hiervoor zijn reeds enige stappen ondernomen waarbij het National Bureau of Standards (NBS) te Washington DC als distributiecentrum zal optreden. Een NATO studiegroep (AC-243, Panel III, RSG-10) heeft reeds een databestand voor aaneengesloten cijfers samengesteld en geëvalueerd.

Als voorbeeld van de evaluatie van een woordherkenner worden in Tabel III de bovengenoemde prestatiematen voor de konditie binnen en buiten de vocabulaire gegeven, voor een herkenner van geïsoleerde woorden, de VECSYS RMI-88. Deze woordherkenner is sprekerafhankelijk, kent één referentiepatroon per woord en is nogal gevoelig voor stoorlawaai. Door het toepassen van een speciale mikrofoon en mikrofoonversterker werd de kombinatie ongevoeliger voor stoorlawaai gemaakt. In de tabel zijn de herkenningspercentages gegeven voor een stoorlawaainiveau en stilte. Als stoorlawaai werd ruis toegepast met een spektrum identiek aan het gemiddelde spraakspektrum. Tevens waren niveaufluktuaties, zoals die optreden bij spraak, aangebracht. Op deze wijze werd de invloed van een storende spreker gesimuleerd.

Tabel III Gemiddelde herkenningspecentages voor vijf sprekers, een vocabulaire van 30 woorden. Er werden een testserie van 30 woorden binnen de vocabulaire en een testserie van 30 woorden buiten de vocabulaire gebruikt. Het herkenningspercentage werd gemeten voor geen omgevingslawaai en een omgevingslawaainiveau van 75 dB(A). De percentages zijn vermeld voor een akseptatiedrempel van de herkenner van resp. 30 en 25 (25 tussen haakjes).

akseptatiedrempel = 30 (25)		stilte	75 dB(A)
binn voc	terecht herkend	94,7% (86,0%)	92,0% (84,0%)
	verkeerd herkend	0,7% (0,0%)	2,0% (1,3%)
	niet herkend	4,7% (14,0%)	6,0% (14,7%)
buit voc	verkeerd herkend	34,0% (5,0%)	36,2% (7,4%)

4. TOEKOMSTVERWACHTINGEN VOOR HET EVALUEREN VAN SPRAAKPRODUKTIE- EN SPRAAKHERKENNINGSSYSTEMEN

Ten aanzien van het testen van spraakproduktiesystemen is goede aansluiting gevonden met methodieken die reeds beschikbaar waren voor het testen van spraakkommunika-

tiekanalen. Immers in beide gevallen gaat het om de kwaliteit van het voortgebrachte spraaksignaal. Het is mogelijk dat op dit gebied enige normalisatie van de toegepaste testmethoden nodig is.

Bij automatische spraakherkenning ligt dit anders. Hier is de ervaring zeer jong en beperkt en zullen de toegepaste methodieken hun waarde in de praktijk nog moeten bewijzen. Toch wordt reeds gewerkt aan een normalisatie. Hiervoor probeert IEEE voorstellen te doen, tevens is een internationaal databestand (NBS) in opbouw en wordt in verschillende internationale lichamen (NATO e.d.) aan standaardisatie en uitwisseling van gegevens gewerkt. Hierbij komen ook ergonomische aspekten (human factors) aan de orde. Het kommercieel beschikbaar komen van spraakherkennende systemen tegen een aanvaardbare prijs zal deze ontwikkeling zeker versnellen.

5. REFERENTIES

[1] Voiers W.D.: Diagnostic Evaluation of Speech Intelligibility. Chap. 32 in M.E. Hawley (ed.). Speech intelligibility and speaker recognition, Vol. 2. Benchmark Papers in Acoustics, Dowden, Hutchinson, and Ross, Stroudsburg, Pa., 1977.

[2] Steeneken H.J.M.: Ontwikkeling en Toetsing van een Nederlandstalige Diagnostische Rijmtest voor het testen van Spraakkommunikatiekanalen. Rapport IZF 1982-13, Instituut voor Zintuigfysiologie TNO, Soesterberg, 1982.

[3] Martin T.B., Welch J.R.: Practical Speech Recognizers and some Performance effectiveness parameters. In Trends in Speech Recognition Research, Prentice Hall, N.J., 1980.

[4] Michael Nye J.: Human Factors Analysis of Speech Recognition Systems. Speech Technology, Vol. 1, No. 2, 1982.

[5] Lea W.A.: Selecting the best Speech Recognizer for the job. Speech Technology, Vol. 1, No. 4, 1983.

Voordracht gehouden tijdens de 319e werkvergadering.

ONTWERPASPECTEN VAN ANALOGE TRANSDUCENTEN ELEKTRONICA DEEL II

RUIS EN OFFSET IN GEÏNTEGREERDE VERSTERKERS

Dr.ir. G.C.M. Meijer

Technische Hogeschool Delft, Afdeling der Elektrotechniek

1. INLEIDING

Bij het versterken van kleine signalen zijn ruis en offset veelal de belangrijkste verschijnselen die de resolutie beperken. De afgelopen tien jaar zijn de versterkers in dit opzicht aanmerkelijk verbeterd. Met name is dit het geval bij operationele versterkers (opamps), waarbij de voornaamste factoren voor de toegenomen kwaliteit de volgende zijn:

- Door een betere beheersing van IC fabricageprocessen is de gelijkheid (matching) van componenten op één chip met sprongen vooruitgegaan. De offset, welke ontstaat door onbalans in de ingangstrap, is hierdoor aanzienlijk gereduceerd.
- Door de komst van trimtechnieken is het mogelijk geworden, om al tijdens de IC produktie de (toch al lage) offset met circa een factor tien te reduceren.
- Door een betere procesbeheersing is het aantal oppervlaktetoestanden en daarmee de 1/f ruis minder geworden.
- Door een verbetering van de hoog-frequentie eigenschappen van de transistoren is ook de ruis bij hogere frequenties afgenomen.
- Door de komst van bifet processen, waarbij men op één chip zowel junctiefet's als bipolaire transistoren maakt, kunnen voor de ingangstransistoren fet's worden toegepast, waardoor bij hoge bronimpedanties een drastische verbetering in zowel het offset- als in het ruisgedrag gerealiseerd kan worden.

In dit artikel wordt het ruis- en offsetgedrag van enige moderne opamps besproken. Vervolgens wordt nagegaan hoe ingangstrappen in dit opzicht geoptimaliseerd kunnen worden en in hoeverre goede offseteigenschappen verenigbaar zijn met goede ruiseigenschappen. Allereerst zullen we beginnen met een inventarisatie van de belangrijkste verschillen en overeenkomsten tussen de verschijnselen offset en ruis.

2. VERSCHILLEN EN OVEREENKOMSTEN IN HET RUIS- EN OFFSETGEDRAG VAN VERSTERKERS

Uit het oogpunt van de elektronische ontwerper ziet men de volgende overeenkomsten tussen ruis en offset in versterkers.

a) Zowel de ruis- als de offseteigenschappen hangen af van de grootte van de bronimpedantie Z_g. Deze afhankelijkheid kan worden weergegeven met een equivalente spannings- en stroombron aan de ingang van de versterker (Fig. 1).

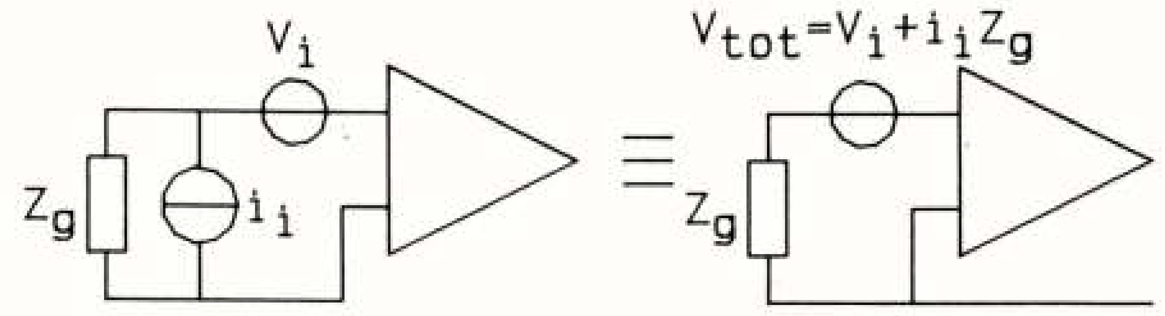


Fig. 1(a) Het ruis- en offsetgedrag van een versterker weergegeven met een equivalente spannings- en stroombron aan de ingang van de versterker.

(b) Alternatief schema.

b) De invloed van het ruis- en offsetgedrag van een versterker kan niet verminderd worden door terugkoppeling toe te passen. Bij een ongunstig gekozen terugkoppelnetwerk kan wel verslechtering van dit gedrag optreden.

c) Bij een goed ontworpen versterker wordt de offset en ruis bepaald door de eerste versterkertrap.

d) Het ruis/offsetgedrag kan geoptimaliseerd worden door geschikte keuzen van de typen componenten en de grootten van de instellingen.

De ontwerper van versterkers wordt geconfronteerd met de volgende verschillen tussen ruis en offset:

e) Offset heeft een lage tot zeer lage frequentie, terwijl ruis in ieder frequentiegebied een rol speelt. Omdat offset temperatuurafhankelijk is kunnen offsetverschijnselen het "aanzien" van laag-frequente ruis krijgen. Figuur 2 toont als voorbeeld de uitgangsspanning van een referentiespanningsbron onder normale laboratoriumomstandigheden.

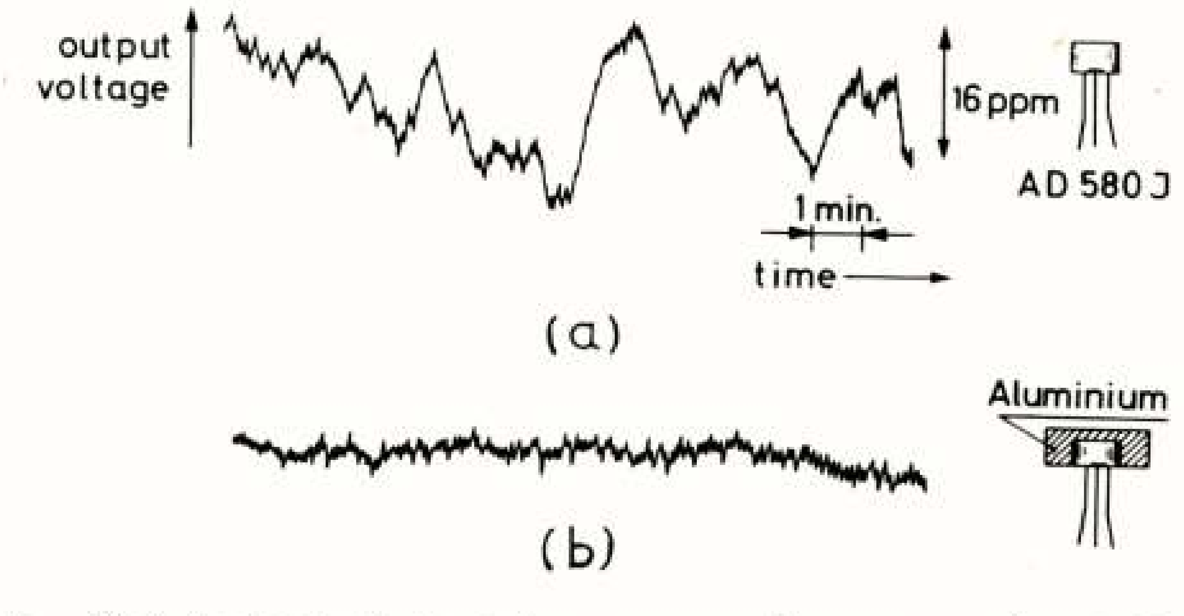


Fig. 2(a) Luchtturbulenties veroorzaken een ruisachtig verloop van de uitgangsspanning van een referentiespanningsbron. (b) Door het aanbrengen van een thermische buffer wordt dit verschijnsel verholpen.

De combinatie van interne dissipatie, afkoeling door luchtturbulentie en de temperatuurafhankelijke offset-spanning in de ingangsversterker veroorzaken de ruisachtige fluctuaties van de uitgangsspanning. Met behulp van een thermische buffer kan dit verschijnsel aanzienlijk gereduceerd worden (Fig. 2(b)).

f) De beste keuze van componenten en instelling voor optimaal ruisgedrag verschilt enigszins van die voor optimaal offsetgedrag.

g) Offset kan gecompenseerd worden; ruis niet. Indien de compensatie van de offset voor honderd procent effectief zou zijn dan zouden we verder alleen maar op de ruiseigenschappen hoeven te letten. Helaas blijkt dit niet het geval te zijn. We zullen nu eerst nagaan hoe deze compensatie gerealiseerd kan worden en hoe effectief die in de praktijk blijkt te zijn.

3. METHODEN OM OFFSET TE TRIMMEN

De offset van een versterker kan gereduceerd worden door de onbalans in de versterkertrappen te verminderen, of door compensatie met speciaal hiervoor aangebrachte stroom- en spanningsbronnen.

De offset wordt afgeregeld met speciaal hiervoor aangebrachte weerstanden in de versterker of in het compensatiecircuit.

Het afregelen van weerstanden kan op een aantal manieren gebeuren:

De IC gebruiker kan dit doen door

- het afregelen van instelpotentiometers; hetgeen duur en arbeidsintensief is,
- het aanbrengen van uitgezochte weerstanden; hetgeen bij massaproduktie geheel automatisch en daardoor goedkoop kan gebeuren.

De IC fabrikant kan afregelen met behulp van

- het kortsluiten van zenerdioden (zener-zapping),
- het doorbranden van verbindingen (fusable-links),
- het lasertrimmen van dunne - filmweerstanden.

Bij de momenteel veel gebruikte zener-zapmethode worden in-seriegeschakelde weerstanden geshunt door zenerdioden (Fig. 3(a)).

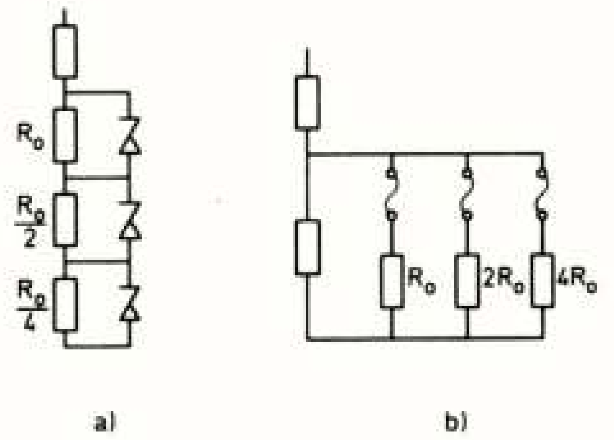


Fig. 3 De IC fabrikant kan weerstanden afregelen door bij voorbeeld (a) zenerdioden kort te sluiten of (b) verbindingen door te branden.

Deze zenerdioden kunnen met een korte stroomstoot worden opgeblazen en vormen dan een zeer betrouwbare kortsluiting met een weerstand van circa 1Ω.

Bij de fusable-link methode staan weerstanden parallel geschakeld (Fig. 3(b)). Door het opblazen van één of meer verbindingen worden deze weerstanden uit het circuit verwijderd.

Bij het zogenaamde laser-trimmen worden met een laserbundel stukjes weerstandsmateriaal uit opgedampte dunne-filmweerstanden "weggebrand".

4. OFFSET EN RUIS IN OPERATIONELE VERSTERKERS

In deze paragraaf zullen we voor een aantal moderne laser- getrimde opamps nagaan in hoeverre goede ruiseigenschappen samengaan met goede ruiseigenschappen en wat bij zeer hoge frequenties nu het grootste probleem is: offset dan wel ruis. De ingangstrappen van de opamps zijn bepalend voor het ruis en offsetgedrag. Voor een vergelijking tussen diverse soorten opamps selecteerden we typen met verschillende soorten ingangscomponenten:

a) Bipolaire transistoren. Deze transistoren hebben een lage offset- en ruisspanning aan de ingang en daarentegen een vrij grote offset en ruisstroom. Hierdoor zijn deze ingangstrappen vooral geschikt bij lage bronimpedanties.

b) Super-bèta transistoren. Deze bijzondere bipolaire transistoren hebben een zeer hoge stroomversterkingsfactor (3000 - 10.000). De basisstroom is daardoor klein waardoor het ruisgedrag bij hoge bronimpedanties relatief gunstig is.

c) Junctiefet's. Deze componenten hebben een te verwaarlozen ingangsstroom en hebben daardoor bij hoge bronimpedanties superieure eigenschappen. Bij lage frequenties neemt de ruisdichtheid (ruis per eenheid van bandbreedte) sterk toe.

De grootte van de ingangsruisspanning V_i in het frequentiegebied $0{,}1Hz < f < 10Hz$ voor de drie geselecteerde opamps is vermeld in tabel I. Tevens is in deze tabel de temperatuurcoëfficient van de ingangsoffsetspanning vermeld.

Tabel I De ingangsoffsetspanning en de laag-frequente ruisspanning voor de diverse opamp's.

	laser getrimde opamp's		
	bipolair AD OP-07	super-bèta AD 517	bifet AD 547
ruis 0,1Hz<f<10Hz	$0{,}35\mu V_{p-p}$	$2\mu V_{p-p}$	$2\mu V_{p-p}$
temperatuur coëfficient v.d. offset spanning	$0{,}3\mu V/^oC$	$1\mu V/^oC$	$2\mu V/^oC$

Het blijkt dat de grootte van de piek-piek waarde van de ruis overeenkomt met de verandering in de offsetspanning bij een temperatuurstijging van circa 1°C. Daar versterkers over het algemeen over een temperatuurgebied van tientallen graden moeten functioneren kan men stellen dat offset een groter probleem is dan laagfrequente (1/f) ruis. Dit blijkt ook te gelden indien men de ingangsoffsetstromen vergelijkt met de 1/f ingangsruisstromen.

De totale offset $|v_i + i_i Z_g|$ van de drie typen versterkers als functie van de bronimpedantie $|Z_g|$ is weergegeven in Fig. 4.

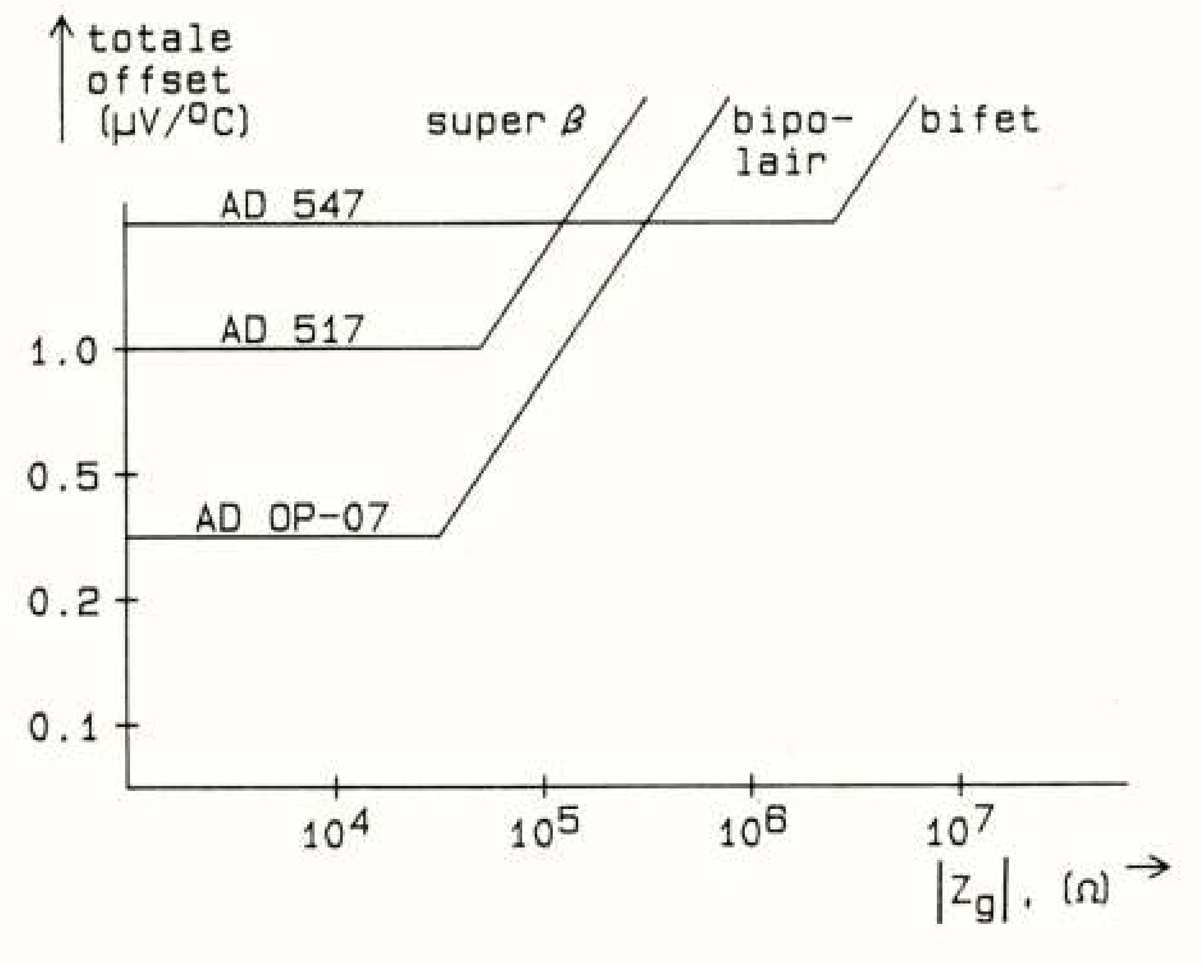


Fig. 4 De totale offset $|v_i + i_i Z_g|$ als functie van de bronimpedantie Z_g.

Bij een bronimpedantie lager dan 200kΩ heeft de bipolaire opamp de beste eigenschappen. Daarboven is de bifet opamp het beste. Het lijkt misschien vreemd dat de superbèta opamp ook bij hoge bronimpedanties slechter is dan de gewone bipolaire opamp. De reden hiervoor is dat de basisstromen (ingangsstromen) van super-bèta transistoren weliswaar erg laag zijn, maar dat de relatieve gelijkheid (matching) ervan zó slecht is dat de offsetstroom, welke ontstaat door het verschil in basisstromen, voor super-bèta opamps nog groter is dan voor "gewone" bipolaire opamps.

Bij frequenties die niet zeer laag zijn speelt offset geen rol. Indien de frequentie hoger is dan de zogenaamde ruishoekfrequentie wordt het $\frac{1}{f}$ ruisgedrag overheerst door witte ruis. Deze ruis is over een vrij groot gebied onafhankelijk van de frequentie. De totale ruisspanning $v_n + i_n Z_g$ aan de ingang voor dit frequentiegebied is in Figuur 5 weergegeven als functie van de bronimpedantie. De gebroken lijn stelt de ruis van een reëele bronimpedantie Z_g voor.

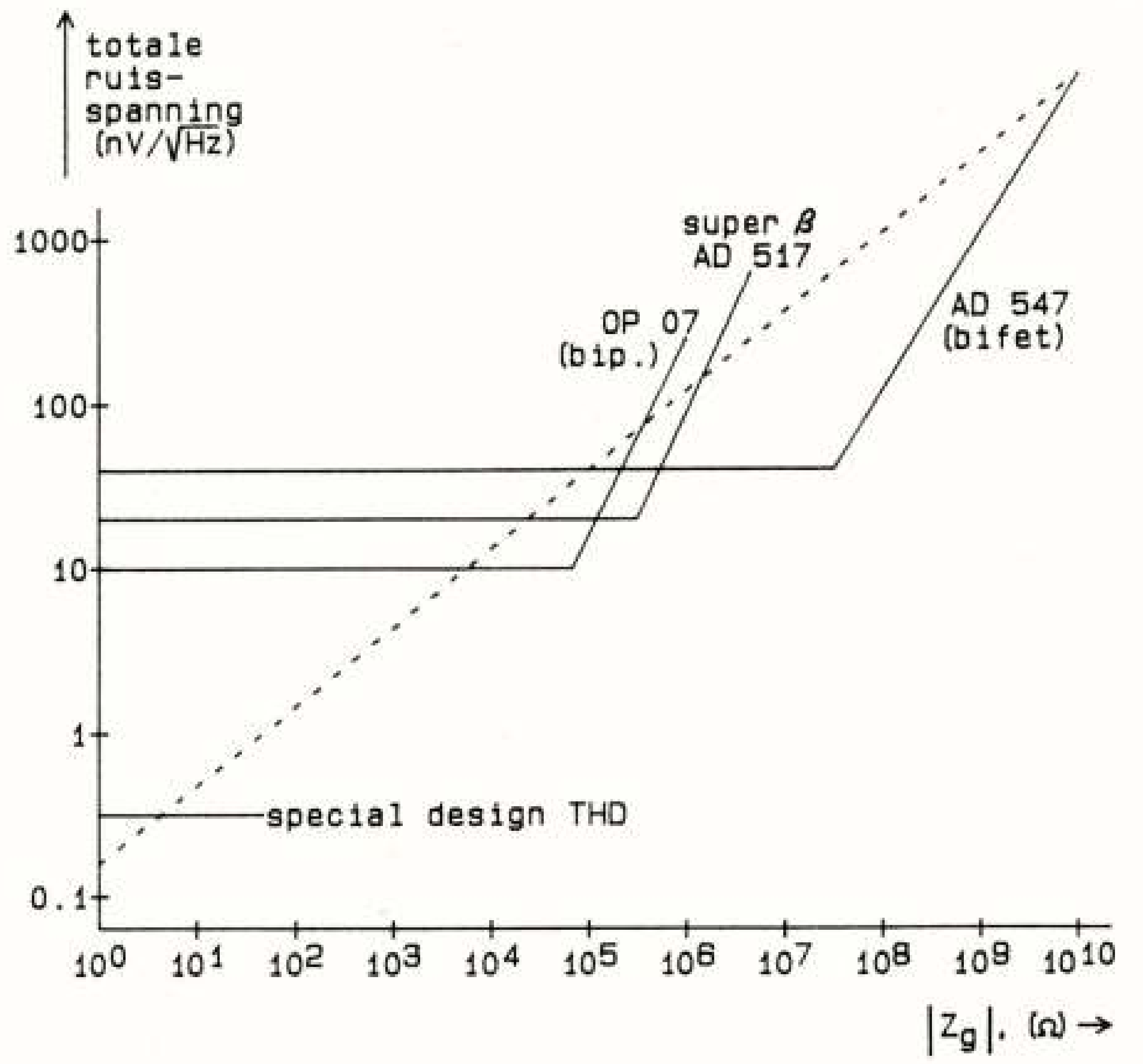


Fig. 5 De totale witte-ruisspanning per eenheid van bandbreedte aan de ingang van de opamp als functie van de bronimpedantie Z_g.

Bij lage bronimpedanties is de bipolaire opamp duidelijk in het voordeel. De ruis van de bipolaire opamp is bij lage bronimpedanties in hoofdzaak afkomstig van de basisweerstand van de ingangstransistoren. Deze weerstand kan geminimaliseerd worden door een groot aantal ingangstransistoren parallel te schakelen of door de ingangstransistoren te voorzien van een aantal parallel geschakelde lange smalle emitters. Dit is toegepast in een -op de TH ontwikkelde- ruisarme versterker, die bij lage bronimpedanties extreem goede ruiseigenschappen heeft (Fig. 5), [1]. Bij hoge bronimpedanties heeft de bifet opamp zeer gunstige eigenschappen.

De super-bèta opamp komt slechts over een zeer klein gebied van bronimpedanties als beste uit de bus. Daar ook de offset eigenschappen nogal te wensen over laten, zijn er op dit moment nog nauwelijks redenen om super-bèta transistoren toe te passen.

5. OPTIMALISATIE VAN DE INGANGSTRAP MET BETREKKING TOT RUIS EN OFFSET

5.1 Indien met in de gelegenheid is om zelf de ingangstrap van een versterker te ontwerpen dan kan men meestal tot een kwalitatief beter produkt komen dan wanneer men is aangewezen op opamps. Bij het ontwerpen van een ingangstrap bepaalt men eerst de bronimpedantie en afhankelijk van de grootte ervan kiest men
- het type van de gebruikte componenten (JFET/bip.)
- de instelstroom.

Afhankelijk van verdere eisen die men aan de versterker stelt (bijvoorbeeld ten aanzien van bandbreedte, doorslagspanning, versterking enz.) maakt men een keuze uit

een aantal mogelijke configuraties voor de ingangstrap. Wij zullen in paragraaf 5.2 voor een aantal gangbare configuraties nagaan welke componenten het meest tot de ruis en offset bijdragen. Daarna wordt in paragraaf 5.3 de keuze van de instelstroom besproken.

5.2 Configuraties voor de ingangstrap

De eenvoudige verschilversterker met belastingsweerstanden (Fig. 6) blijkt vrij gunstige ruis- en offseteigenschappen te hebben en wordt daarom veel toegepast. De thermische ruis van de weerstand kan gemodelleerd worden met een equivalente stroombron ter grootte van $(4kT\,\Delta f/R)^{\frac{1}{2}}$ parallel aan de weerstand, waarbij Δf de ruisbandbreedte voorstelt (Fig. 6(b)).

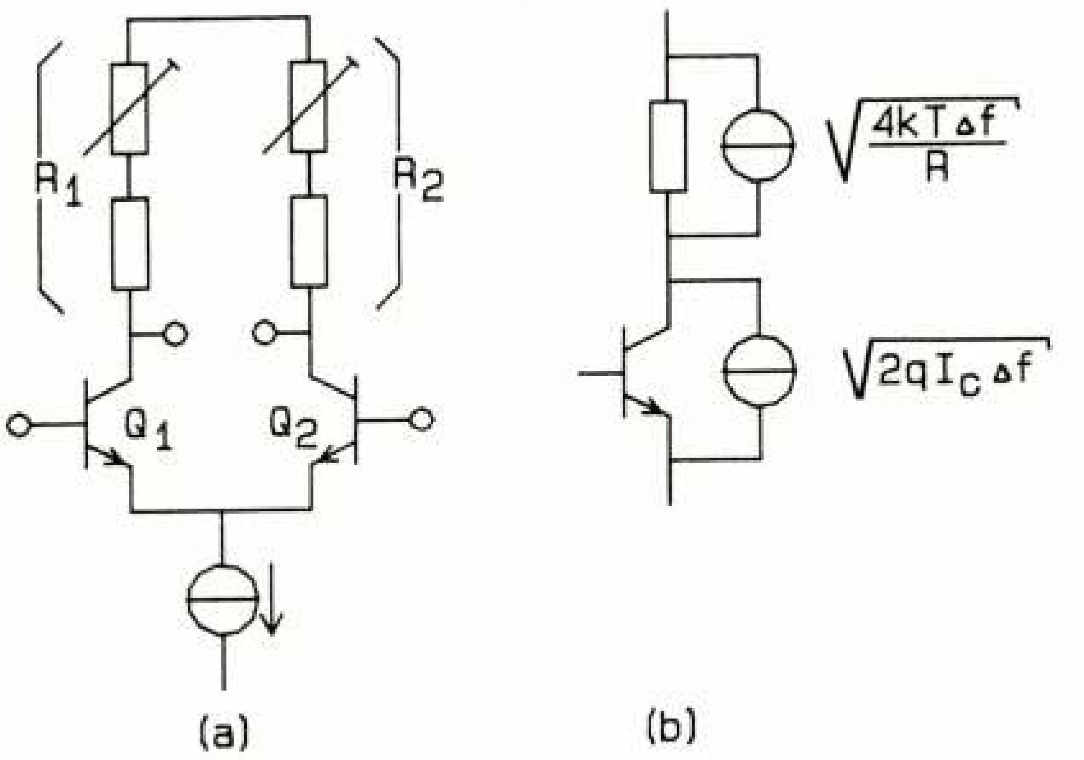


Fig. 6 (a) Verschilversterkertrap met ohmse belastingsweerstanden. (b) Model voor de thermische ruis van de weerstand en de hagelruis van de transistor.

De hagelruis van de transistor geven we weer met een stroombron $(2qI_C\Delta f)^{\frac{1}{2}}$, waarin I_C de instelstroom van de transistor is.

De thermische ruis van de weerstand is verwaarloosbaar t.o.v. de hagelruis van de transistor als geldt dat

$$4kT/R \ll 2qI_C \qquad (1)$$

ofwel bij T = 300K

$$I_C R \gg 2\frac{kT}{q} \approx 50\text{mV} \qquad (2)$$

Vergelijking (2) verschaft ons een praktische ontwerpregel: Indien serieweerstanden zijn opgenomen in de collector- of emitterketen van bipolaire transistoren dan kan men de thermische ruis van de weerstanden verwaarlozen als de gelijkspanning over die weerstanden veel groter is dan $2(kT/q) \approx 50$mV.

De offset van deze trap is een gevolg van onbalans die onder andere ontstaat door ongelijkheid (mismatching) tussen de gebruikte componenten. Daar de matching van weerstanden beter is dan van transistoren kan men stellen dat de transistoren de voornaamste bron voor zowel de ruis als de offset vormen. Om de gevolgen van het Early effect te verminderen worden de ingangstransistoren bijna altijd gecascodeerd (Fig. 7). Men kan gemakkelijk aantonen dat deze cascodetransistoren nauwelijks invloed hebben op het ruis- en offsetgedrag.

Bij kleine instelstromen dienen de belastingsweerstanden een vrij hoge waarde te hebben om voldoende versterking te realiseren. Omdat het in de IC techniek vaak moeilijk is om hoogohmige weerstanden te maken past men daarom vaak aktieve belastingen toe

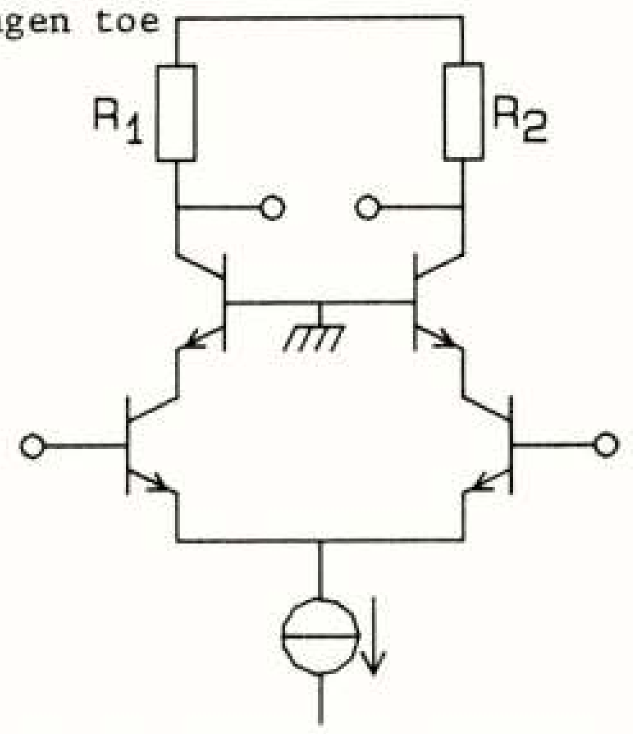


Fig. 7 Om de gevolgen van het Early effect te verminderen worden de ingangstransistoren gecascodeerd.

Bij kleine instelstromen dienen de belastingsweerstanden een vrij hoge waarde te hebben om voldoende versterking te realiseren. Omdat het in de IC techniek vaak moeilijk is om hoogohmige weerstanden te maken past men daarom vaak aktieve belastingen toe (Fig. 8)

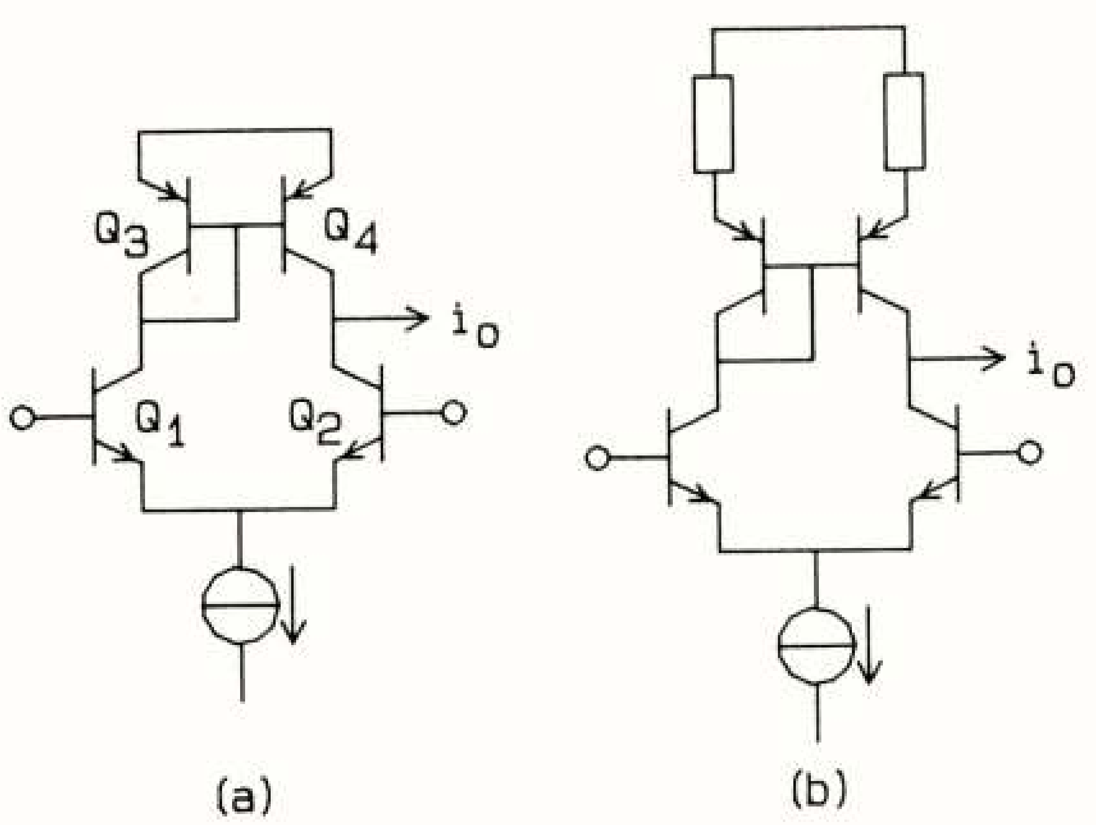


Fig. 8 (a) Een ingangstrap met een stroomspiegel als belasting. (b) Door de stroomspiegeltransistoren te voorzien van emitterweerstanden worden de ruis- en offseteigenschappen beter.

Indien de bronimpedantie laag is dan zijn de transistoren Q_1 en Q_2 ongeveer op dezelfde wijze in de schakeling opgenomen als de spiegeltransistoren Q_3 en Q_4. Daardoor dragen de vier transistoren Q_1 - Q_4 ongeveer evenveel bij tot de ruis en de offset van de trap. Deze

schakeling is daardoor in dit opzicht slechter dan die van Fig. 6.

Men kan de schakeling verbeteren door de stroomspiegeltransistoren te voorzien van emitterweerstanden (Fig. 8(b). Door de hierdoor verkregen tegenkoppeling in het stroomspiegelcircuit vermindert de ruisbijdrage van Q_3 en Q_4 ongeveer met een factor $(I_C R)/(kT/q)$, waarbij $I_C R$ de gelijkspanning over de weerstanden is.

De ruisbijdrage van de stroomspiegel kan dus gemakkelijk kleiner gemaakt worden dan die van Q_1 en Q_2. Indien de weerstanden groot genoeg zijn dan zal ook de offset minder worden omdat weerstanden een betere "matching" hebben dan transistoren.

Een andere ingangstrap, die wel in opamps wordt toegepast vanwege de gunstige "level shifting" is die van Fig. 9. In deze configuratie dragen maar liefst zes

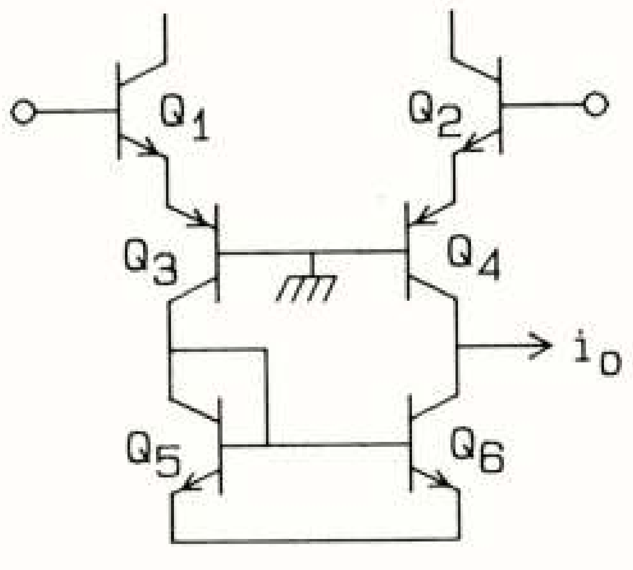


Fig. 9 Ingangstrap met "level shifting" naar de negatieve voedingsspanning.

transistoren bij tot de ruis en de offset. Bij lage bronimpedanties leveren deze transistoren evengrote bijdragen. Deze configuratie is daarom af te raden indien men prijs stelt op goede ruis- en offseteigenschappen.

Indien FET's in de ingangstrappen worden toegepast, dan kan men een soortgelijke afweging maken om de belangrijkste ruisbronnen op te sporen. Een FET ruist als een weerstand ter grootte van $\frac{2}{3g_m}$ zodat de ruis van een weerstand die in serie met de drain of de source staat verwaarloosd kan worden als geldt dat $R \gg \frac{2}{3}\frac{1}{g_m}$.

Tot nu toe hebben we onze beschouwingen beperkt tot lage bronimpedanties. Bij hoge bronimpedanties is de equivalente ruis- en offsetstroombron aan de ingang van belang. Deze stroombronnen worden in alle besproken configuraties bepaald door de ingangstransistoren.

5.3 Optimalisatie van de instelstroom

Minimalisatie van de ruis leidt tot een enigszins andere keuze van de instelstroom dan minimalisatie van de offset. Voor de equivalente ruisspanning v_n aan de ingang geldt voor bipolaire transistoren dat

$$v_n^2 = 4kT\,\Delta f\,\left(r_b + \frac{kT}{qI_C}\right) \qquad (3)$$

waarin r_b de basisbulkweerstand is. Deze ruisspanning kan klein gemaakt worden door de collectorstroom I_C voldoende groot te kiezen en door transistoren met een lage waarde van r_b te kiezen.

Voor de equivalente ruisstroom i_n aan de ingang geldt voor bipolaire transistoren dat

$$i_n^2 = 2qI_B\,\Delta f \qquad (4)$$

Bij een bepaalde bronweerstand R_G is de totale ruisbijdrage in $R_g + v_n$ minimaal als geldt dat

$$g_m = \frac{qI_C}{kT} = \frac{\sqrt{h_{FE}}}{R_G + r_b}\,, \qquad (5)$$

waarin h_{FE} de stroomversterkingsfactor in de gemeenschappelijke emitterschakeling is.

De offsetspanning aan de ingang is in eerste benadering onafhankelijk van de instelstroom. De offsetstroom, die ontstaat door ongelijkheid van de basisstromen zal echter kleiner zijn naarmate de instelstroom kleiner is. Een lage instelstroom is dus gunstig om een lage offset te bereiken.

In een aantal moderne opamps (b.v. de OP-07) compenseert men de ingangsbasisstroom door een evengrote maar tegengestelde stroom aan de ingang toe te voeren. Het zal duidelijk zijn dat deze compensatie niet voor ruis optreedt en dat de ingangsruisstroom juist groter wordt door deze maatregel. Indien de offsetstroom van geen belang is (b.v. omdat de signaalfrequenties voldoende hoog zijn) dan kan men deze maatregel beter achterwege laten.

Indien voor de ingangstransistoren FET's worden toegepast dan geldt bij niet te hoge frequenties dat de ruisspanning aan de ingang evenredig is met $(I_D)^{\frac{1}{4}}$ en dat de offsetspanning slechts weinig afhangt van de drainstroom.

Indien de versterkingsfactor van de eerste trap klein is, dan dient ook de invloed van de volgende trappen in rekening te worden gebracht. Deze situatie kan zich vooral bij hogere frequenties gemakkelijk voordoen. Voor hogere frequenties zijn de beschouwingen tamelijk complex en gaarne verwijs ik hiervoor naar de desbetreffende literatuur [2], [3].

5. CONCLUSIES

- De offset van moderne geïntegreerde versterkers wordt aanzienlijk gereduceerd door trimming van de onbalans en door compensatie van de ingangsstromen. Toch is offset bij lage frequenties nog steeds de belangrijkste beperkende factor in de resolutie van de versterkers.

- Bij lage bronimpedanties kan men het beste voor bipolaire opamps kiezen. Bij hoge bronimpedanties zijn de bifet opamps superieur. De super-bèta opamps bieden nauwelijks enig voordeel.
- Met speciaal ontworpen ingangstrappen kunnen vaak aanzienlijk betere ruis- en offseteigenschappen verkregen worden dan met opamps.
- Bij goed ontworpen versterkers worden ruis en offset in hoofdzaak door de ingangstransistoren bepaald.
- De optimale keuze van de instelstroom is voor offset vaak lager dan voor ruis. Eisen ten aanzien van de bandbreedte leiden tot een relatief hoge instelstroom.

REFERENTIES

[1] E.H. Nordholt en W.G.M. Straver, "A single-chip preamplifier for moving-coil phone cartridges", Proceedings of the 1981 Custom Integrated Circuits Conference, Rochester, 1981.

[2] E.H. Nordholt, "Design of high-performance negative-feedback amplifiers", Elsevier Sc. Publ. Comp., Amsterdam, 1983.

[3] P.R. Gray en R.G. Meyer, "Analysis and design of analog integrated circuits", New York, Wiley, 1977

Voordracht gehouden tijdens de 316e werkvergadering.

UIT HET NERG

LEDENMUTATIES

Voorgestelde leden

Ir. H.J. Simons, Diezerplein 28 B, Zwolle.

Nieuwe leden

Ir. D. de Vries, Van Hasseltlaan 380, Delft.

Nieuwe adressen van leden

W.R.M. Arnoldussen, Taniaburg 13, Leeuwarden.
Ir. J.H.L. van der Bij, Oude Convent 22, Weesp.
Ir. W. van Eck, Beekforel 37, Leiden.
C.G.M. van 't Klooster, Driemasterwerf 9, Zoetermeer.
Ir. R.E.M. Matthews, Halleylaan 16, Bilthoven.
F.J. Soede, LE CONSUL 3ème étage, 25 Places des Halles, 67000 Strasbourg, Frankrijk.

Overleden

Ing. J.M. Luyten, Heerbaan 52, Heel.

Werkbezoek Philips Nijmegen - Juni 1984

Het ligt in de bedoeling eind juni een werkbezoek te brengen aan Philips Nijmegen. Hier vindt IC ontwikkeling- en produktie plaats.

Degene die hiervoor belangstelling heeft, kan zich opgeven bij ondergetekende. Men wordt verzocht behalve naam, adres en telefoonnummer, ook op te geven de naam van de huidige werkgever en de eigen specifieke belangstelling.

Met specifieke wensen zal zo mogelijk rekening worden gehouden. Aan het werkbezoek kunnen 20 NERG-leden deelnemen. De definitieve datum en toelating tot deelname zullen zo spoedig mogelijk aan U bekend worden gemaakt.

Dr.Ir.A.J. Vinck
Afd.Elektrotechniek THE
Postbus 513
5600MB Eindhoven

MAATSCHAPPIJ EN TECHNIEK

SYMPOSIUM DE KWETSBAARHEID VAN DE STAD

Op 18 mei a.s. wordt door de Stichting Toekomstbeeld der Techniek een symposium onder de titel "De kwetsbaarheid van de stad" ter gelegenheid van het gereedkomen van de gelijknamige publikatie.

In de zomer van 1977 viel de elektriciteitsvoorziening uit in New York. Pas na 25 uur keerde de spanning terug. Door ontregeling van het autoverkeer had de brandweer moeite branden te bereiken. De beurs van New York en de banken bleven een dag gesloten. De voedselvoorzieningen werd verstoord omdat liften en diepvriezers niet meer werkten en vrachtwagens gebrek aan brandstof kregen.

Zeldzame verstoringen als natuurrampen, verontreiningingen in het drinkwater en stakingen onderstrepen onze afhankelijkheid van infrastructuren, waarvan we het functioneren als vanzelfsprekend ervaren.

Vier technische infrastructuren werden geanalyseerd, nl. water, elektriciteit, gas en telefonie.

Aan de hand van risico-analyse, de theorie van menselijke fouten en crisisbesluitvorming en een aantal praktijkgevallen wordt een schets gegeven van de belangrijkste problemen.

Hieruit resulteren mogelijkheden om de gevolgen te verzachten of te voorkomen. De aanbevelingen zijn zowel van bestuurlijke als van technische aard.

Het programma omvat de volgende lezingen:

Hoe kwetsbaar is de stad?
Prof.ir. H.Wiggerts, TH Delft
Waarborgen voor een ongestoorde drinkwatervoorziening.
ir. Th.G. Martijn, VEWIN
Bestrijding en preventie van ernstige verstoringen.
ir. A.P. Oele, Commissaris der Koningin in Drente
Bestuurlijke besluitvorming bij ernstige verstoringen.
Prof.dr. U. Rosenthal, Erasmus Universiteit,Rotterdam.

Algemene gegevens:

datum:	vrijdag 18 mei 1984, 13.14 uur
plaats:	De Doelen te Rotterdam, kleine zaal
kosten:	100,-- (incl.STT-publicatie nr. 39) studenten f 15,-- (geen publikatie)
betaling:	voor 11 mei 1984 op gironummer 177.70.70 t.n.v. Congresbureau KIvI te 'sGravenhage, onder vermelding van "STT-symposium" en de naam van de deelnemer. Bij verhindering, mits opgegeven voor 11 mei 1984 is terugbetaling mogelijk.
inlichtingen:	Agaath van der Kamp (STT) Postbus 30424, 2500GK 'sGravenhage tel.no. 070-64.68.00 tst. 55 telex 33641

PATO-CURSUS ELECTROMAGNETISCHE COMPATIBILITEIT (EMC)

Onder auspiciën van het Orgaan voor Postacademisch Onderwijs in de Technische Wetenschappen (PATO) wordt op het gebied van Electromagnetische Compatibiliteit (EMC) een 5-daagse cursus georganiseerd aan de Technische Hogeschool Eindhoven. De cursus wordt in twee gedeelten gegeven en wel op 14, 15 en 16 november 1984 en op 22 en 23 november 1984.

De toename van de dichtheid en complexiteit van elektrische en elektronische systemen maakt hedentendage de kans op ongewenste wederzijdse beinvloeding dermate groot dat preventie van deze beinvloeding noodzakelijk is. Dit laatste zowel in het ontwerpstadium van een systeem als bij de definitieve vormgeving, de installatie en het gebruik. Het vakgebied Electromagnetische Compatibiliteit bestudeert de genoemde ongewenste beinvloeding ofwel stoorproblematiek.

Gezien de noodzakelijke preventie dient EMC een geintegreerd onderdeel van het onderwijs in de Electrotechniek te worden. Deze PATO-cursus EMC geeft daartoe een duidelijke aanzet. De cursus schenkt onder meer aandacht aan: EM-veld theorie, karakterisering van het EM-milieu, overspraak, analysemethode stoorproblemen, emissie van een susceptibiliteit voor stoor-signalen aarding en afscherming, stooronderdrukkende technieken en ontwerpcriteria van systemen.

Belangstellenden kunnen nu reeds naam en adres opgeven aan het PATO-bureau, Prinsessegracht 23, Postbus 30424, 2500GK 's-Gravenhage, tel. 070-644957. Zodra de inschrijfformulieren gereed zijn ontvangen zij een exemplaar, met daarbij informatie over het cursusbedrag en het definitieve cursusprogramma.

Conference on Precision Electromagnetic Measurements-CPEM 84.

The 1984 CPEM - the world's leading international biennial conference for electromagnetic metrology and related fundamental constants - will be held on 20-24 August 1984, at Delft University of Technology, The Netherlands.

Over 100 papers will be delivered in sessions entitled:

EM-based fundamental constants & standards
direct current & low frequency
time & frequency
antennes & fields
microwaves & millimeter waves
(micro) computer-aided meausrements, converters
infrared, visible & ultraviolet radiation
electro optics, fiber optics
lasers
cryo-electronics
technical calibration services

Further information from:

Mrs. IJ. Smits
Department of Electrical Engineering
Delft University of Technology
P.O. Box 5031
2600GA Delft, The Netherlands
Telephone: 31 15 781736
Telex: 38151

Conferentie aankondigingen.

Communications '84. 16-18 May in The Birmingham Metropole Hotel. *)

IMACS TC-1 Modelling and Simulation of Electrical Machines and Converters. 17-18 May 1984 Liége (Belgium) Contact adres: Association des ingenieurs electriciens sortis de l'institut Montefiore, Rue Saint-Gilles 31, B-4000, Liege, Belgium.

2nd International conference on optical fiber sensors, september 5-7, 1984 Stuttgart. Contact adres: R.Kist, Fraunhofer-institut, IPM, Heidenhofstrasse 8, D-7800 Freiburg. Tel. 49(761)84081.

10th European conference on optical communication, september 3-6, 1984 Stuttgart. Contact adres: K.Hess, SEL-Research Centre, Hellmuth-Hirth-Str.42, D-7000 Stuttgart 40, Tel. +711/821 5553

International Conference on Road Traffic Data Collection, 5-7 December 1984, IEE, Savoy Place, London WC2, U.K. *)

International conference on Advances in Command Control and Communication Systems, Theory and Applications 16-18 april 1985, Bournemouth International Conference Centre U.K.*)

Third International Conference on Development in Power-System Protection, 17-19 April 1985, London Savoy Place WC2, U.K.*)

*) Contact adres:
Conference Services Department
The Institution of Electrical Engineers
Savoy Place
London WC2R OBL
Tel. 01-240 1871 (Ext 222)

Tijdschrift van het Nederlands Elektronica- en Radiogenootschap

deel 49 - nr. 2 - 1984

Inhoud

Druk: de Witte Eindhoven